Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.2: Aligned Representations

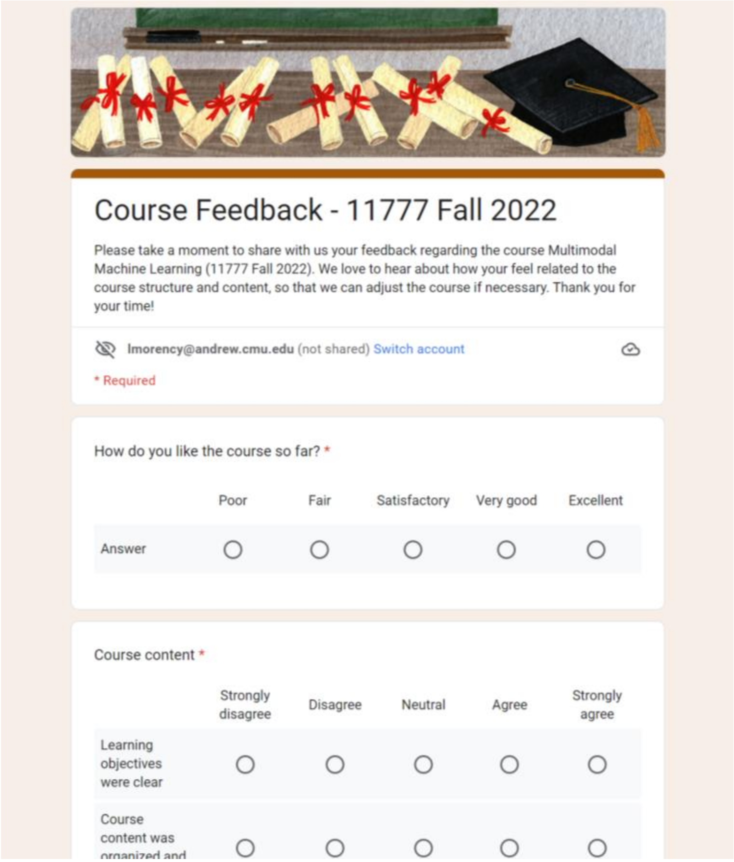
Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Share Your Thoughts!

https://forms.gle/8vmWa7PxBfkGC2i69



Course Feedback - 11777 Fall 2022

Please take a moment to share with us your feedback regarding the course Multimodal Machine Learning (11777 Fall 2022). We love to hear about how your feel related to the course structure and content, so that we can adjust the course if necessary. Thank you for your time!

lmorency@andrew.cmu.edu (not shared) Switch account

* Required

How do you like the course so far? *

	Poor	Fair	Satisfactory	Very good	Excellent
Answer	○	○	○	○	○

Course content *

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Learning objectives were clear	○	○	○	○	○
Course content was organized and	○	○	○	○	○

Deadline
Please submit your feedback about this course before this Wednesday 10/5

Optional,
but greatly appreciated! ☺

Anonymous, by default.

- You can optionally share your email address if you want us to follow-up with you directly.

Second Project Assignment (Due Monday 10/10)

Main goals:

1. Help clarify and expand your research ideas
   - Build qualitative intuitions by directly studying the original data
   - Perform analyses on your dataset, relevant to your research ideas
2. Understand the structure in your data and modalities
   - Perform analyses and visualizations to understand each modality
   - Study representations from CNNs, word2vec, BERT, …

Two types of analyses:

- Idea-oriented analyses
- Modality-oriented analyses

Second Project Assignment (Due Monday 10/10)

Examples of **idea-oriented** analyses:

- What external knowledge is needed when performing the task?
- How often multimodal information is needed? How is it integrated?
- What biases may be present in the data? Which modalities?

Examples of **modality-oriented** analyses:

- What are the different verbs used in the VQA questions?
- What objects do not get detected? Are they important?
- Visualize face embeddings with respect of emotion labels

Second Project Assignment (Due Monday 10/10)

Idea-oriented analyses:

- **Human simulations:** Instead of a computer, try to do the same task as a human. Gather notes on how you perform the task.
- **Data analysis:** study the multimodal data (e.g., using statistical methods) to clarify your hypotheses related to your research ideas

Modality oriented analyses:

- **Language modality:** explore the language structure in your dataset. You can compare word-level and sentence-level embeddings.
- **Visual modality:** study visual representations for your dataset. You visualize how your visual features successfully model your labels.

Second Project Assignment (Due Monday 10/10)

Number of analyses:

- Teams of 3 or 4 students: 2 analyses (4 pages)
- Teams of 5 or 6 students: 3 analyses (6 pages)

- ➢ You can mix and match between idea-oriented and modality-oriented
- ➢ Be sure to talk with your TA about formalizing your analysis plan
- ➢ Each analysis need a separate discussion section

Detailed instructions on Piazza (Resources section)

Multimodal Machine Learning

Lecture 5.2: Aligned Representations

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Objectives of today's class

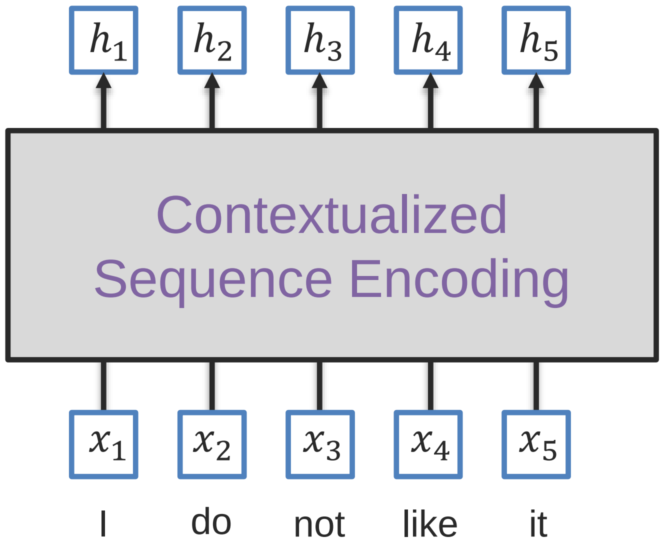
- Contextualized sequence representations
- Transformer networks
  - Self-attention
  - Multi-head attention
  - Position embeddings
  - Sequence-to-sequence modeling
- Multimodal contextualized embeddings
- Language pre-training
  - BERT pre-training and fine-tuning

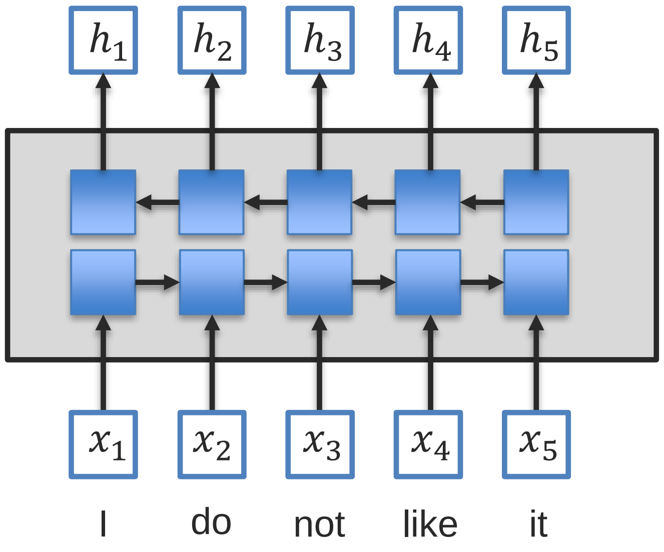
Contextualized Sequence Representations

Sequence Encoding - Contextualization

Option 1: Bi-directional LSTM:

(e.g., ELMO)

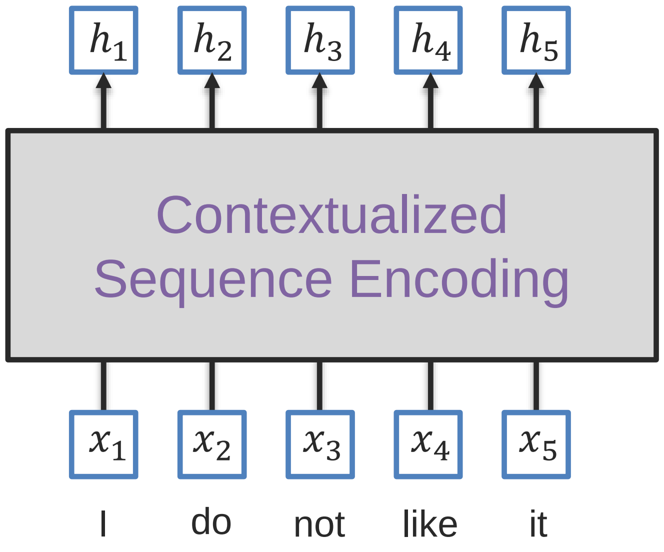




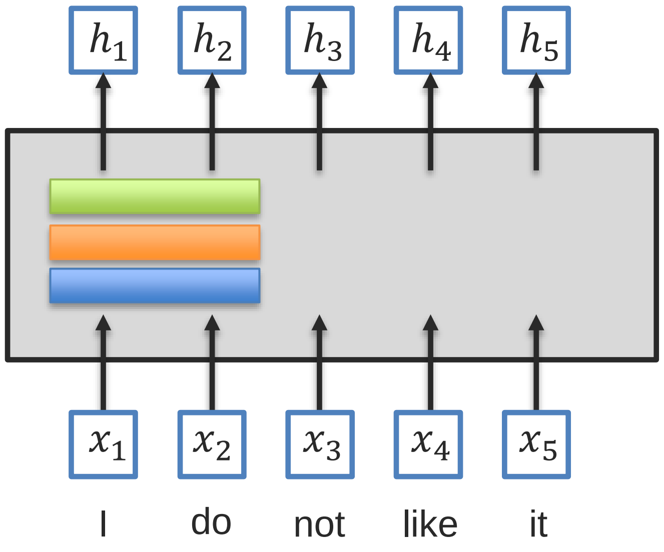
How to encode this sequence while modeling the interaction between elements (e.g., words)?

But harder to parallelize…

Sequence Encoding - Contextualization



Option 2: Convolutions

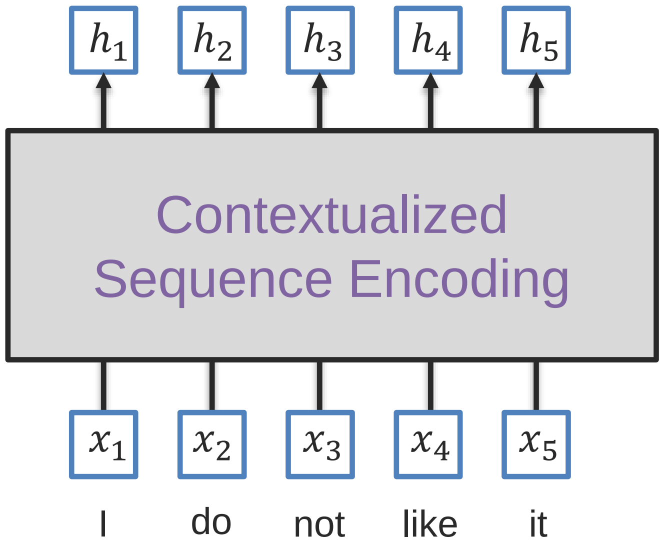


Can be parallelized!

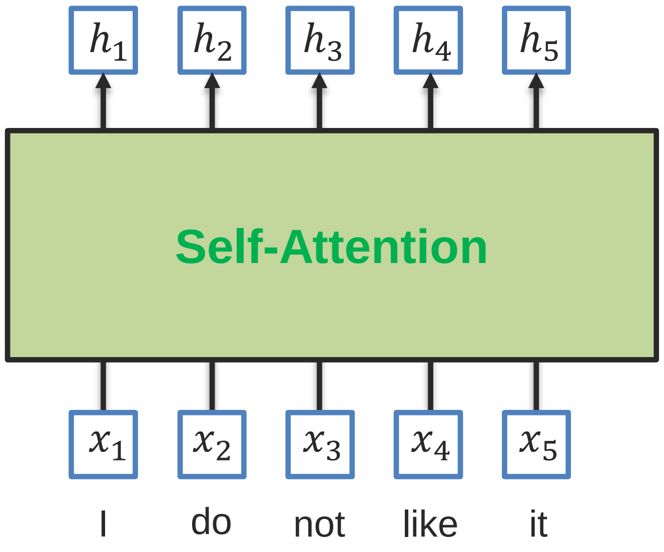
But modeling long-range dependencies require multiple layers

And convolutional kernels are static

Sequence Encoding - Contextualization



Option 3: Self-attention



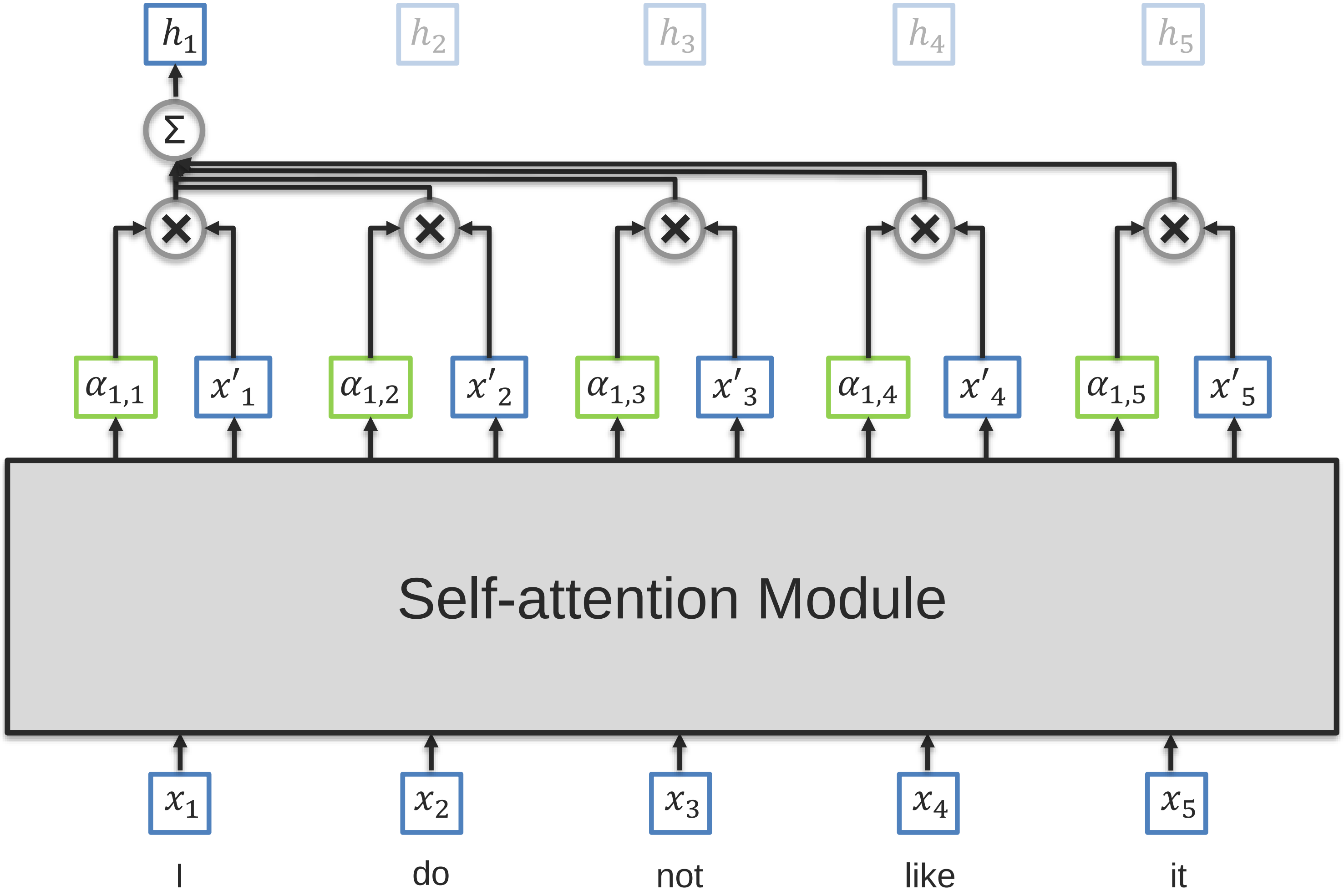
Can be parallelized!

Long-range dependencies

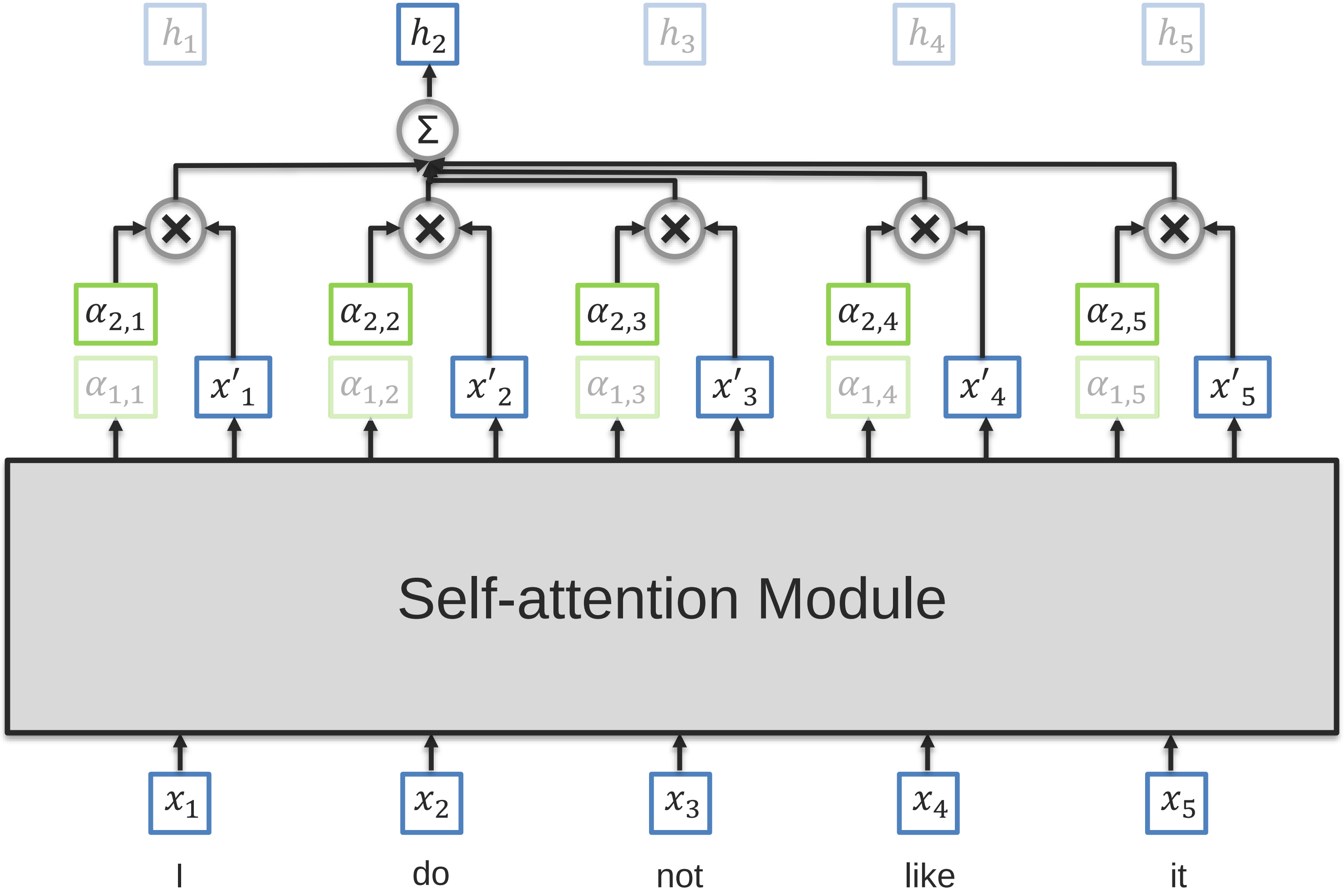
Dynamic attention weights

Self-Attention

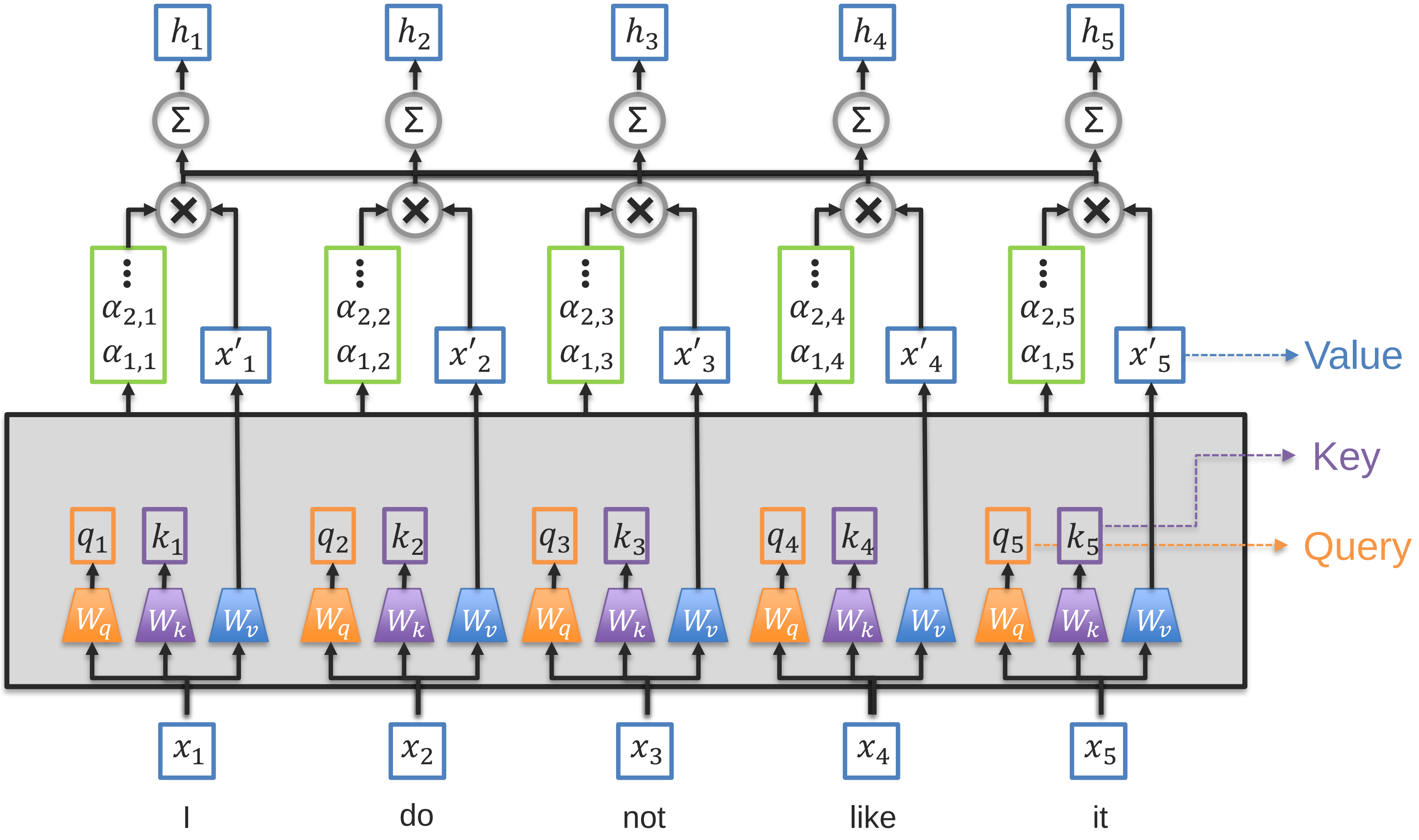
Self-Attention



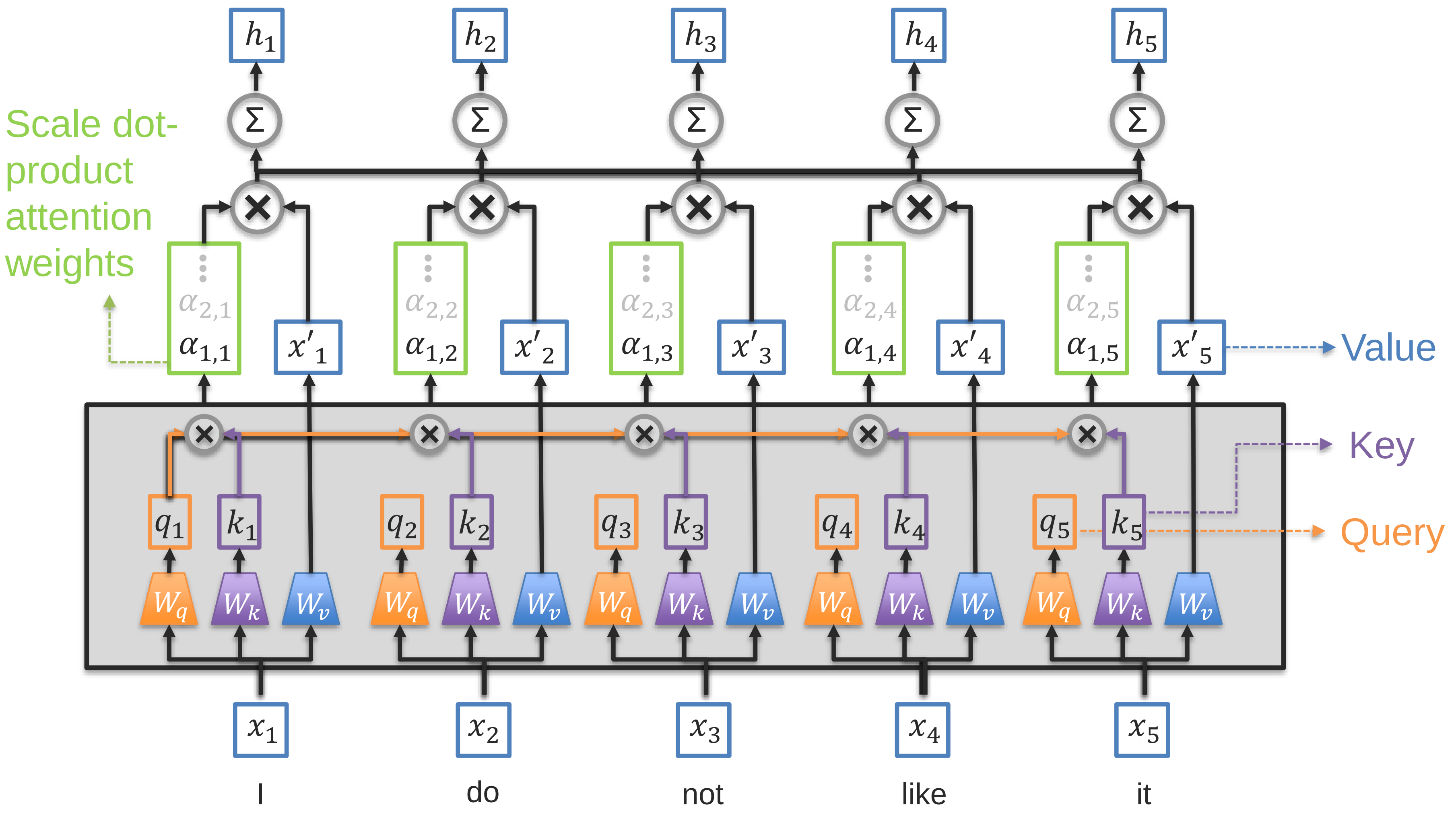
Self-Attention



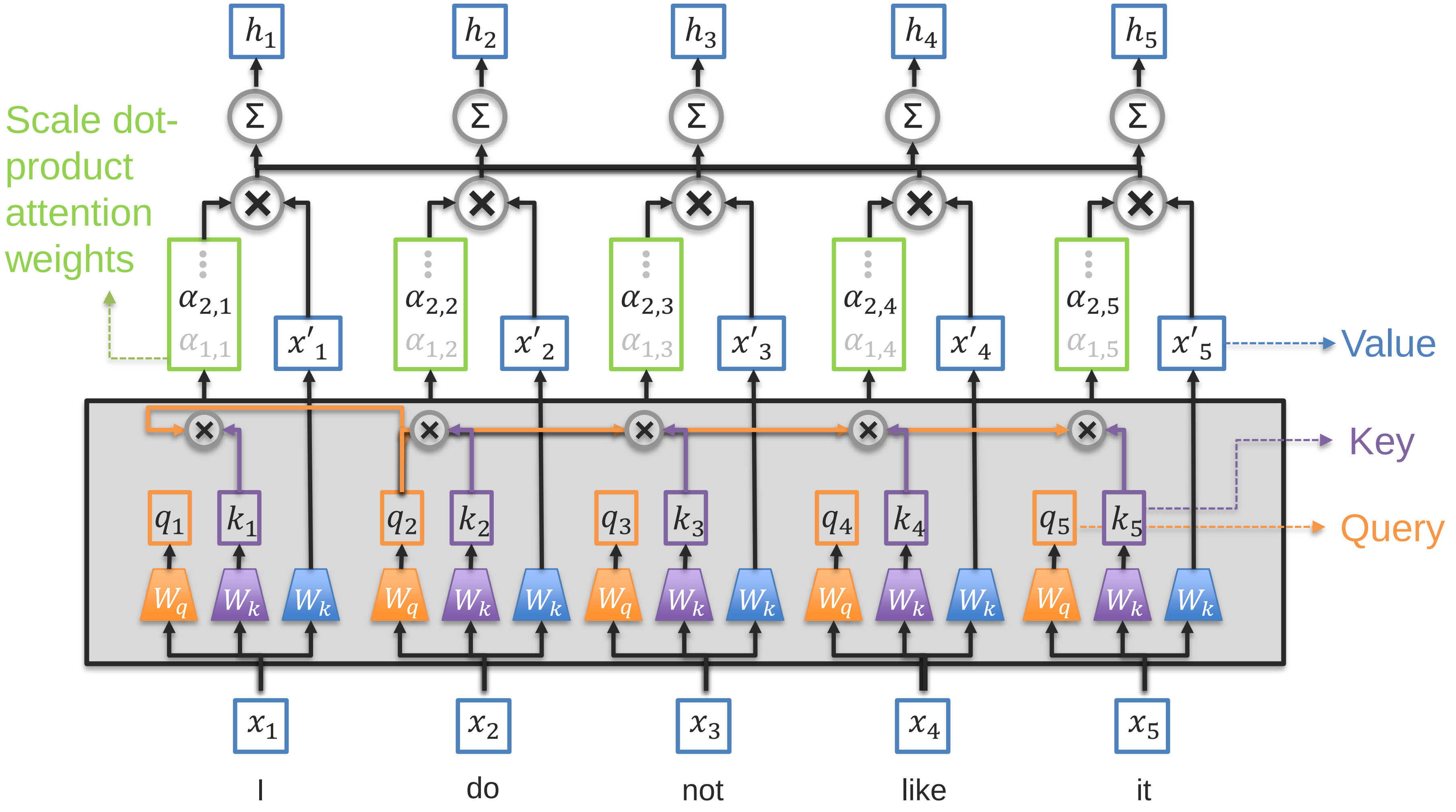
Transformer Self-Attention



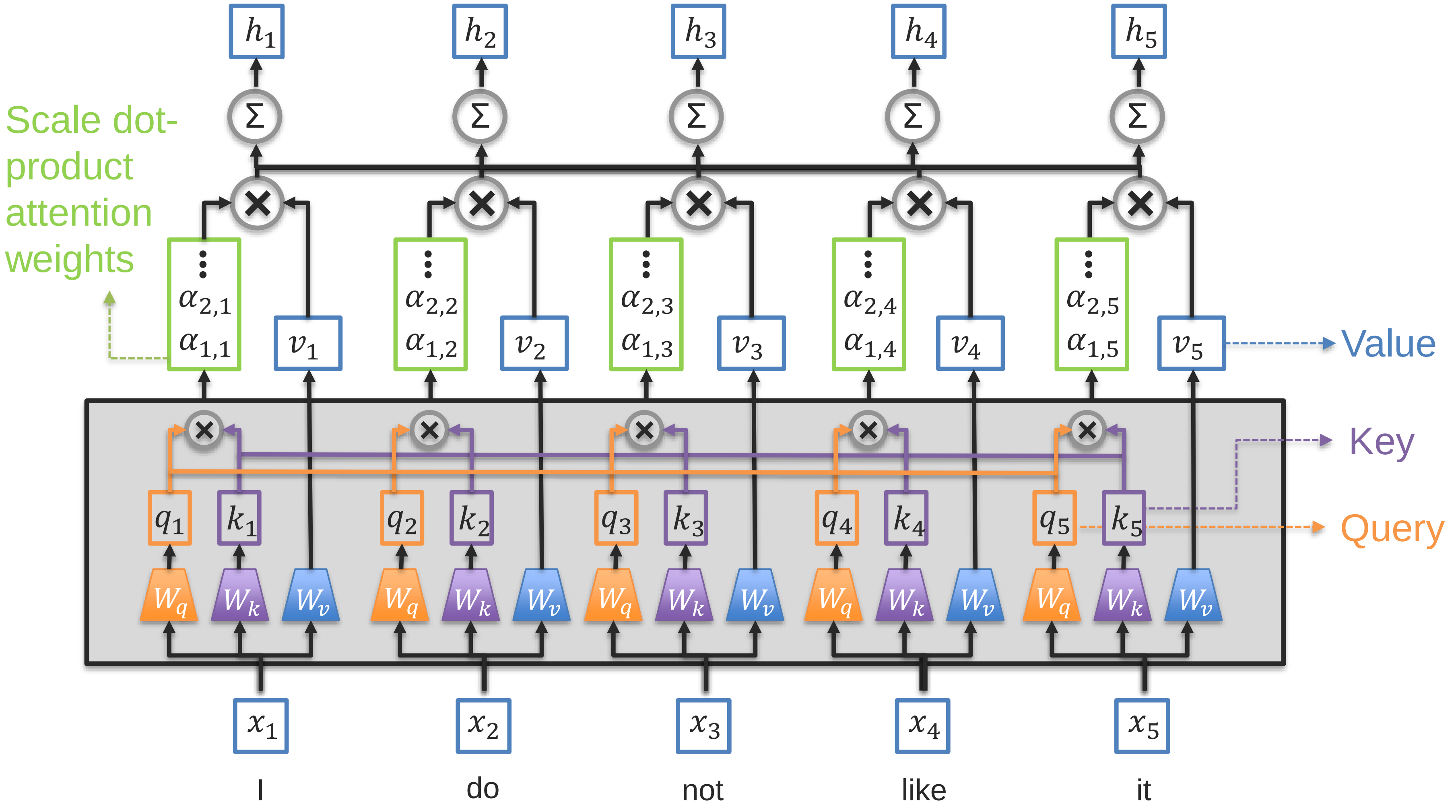
Transformer Self-Attention



Transformer Self-Attention

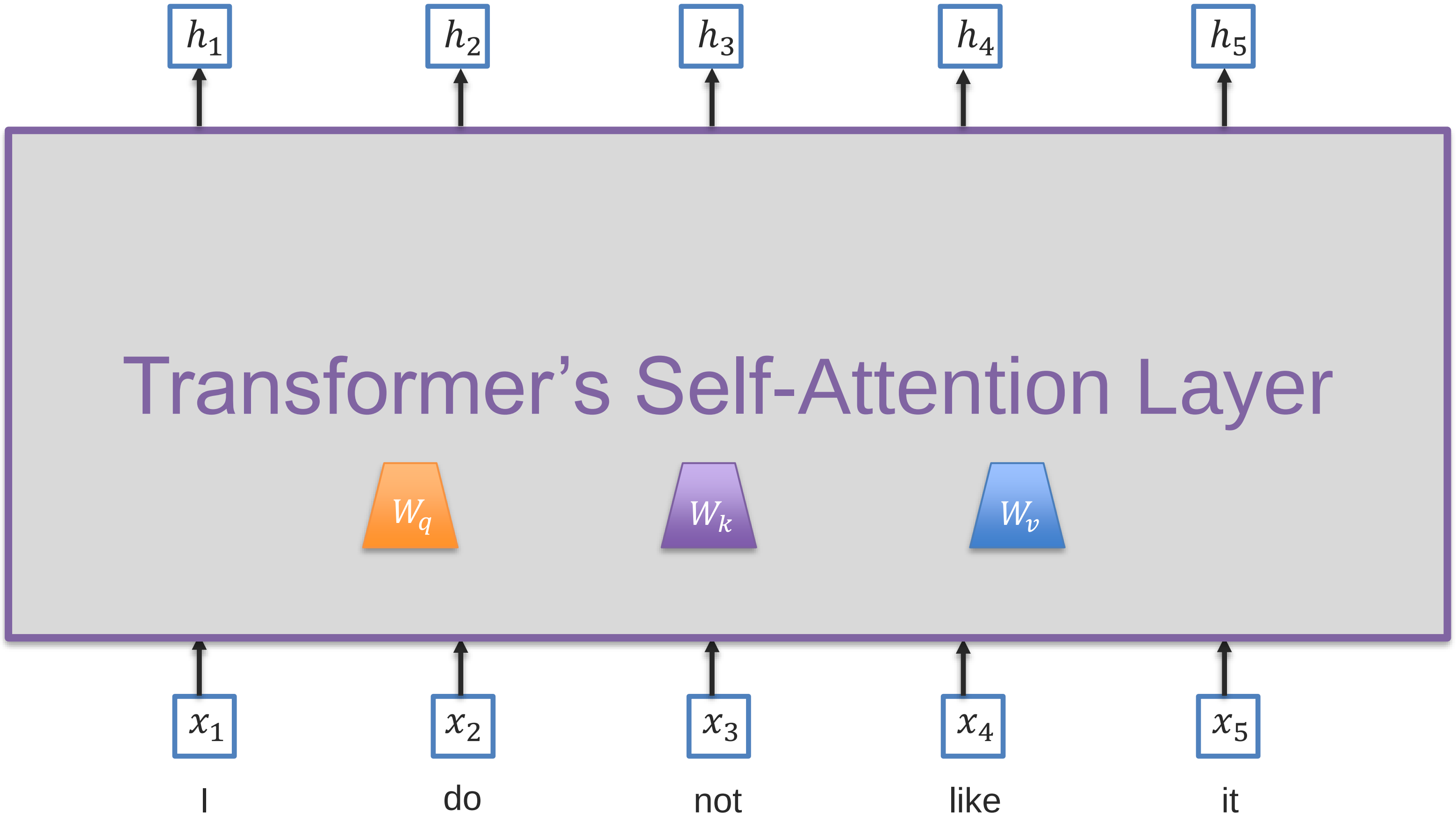


Transformer Self-Attention

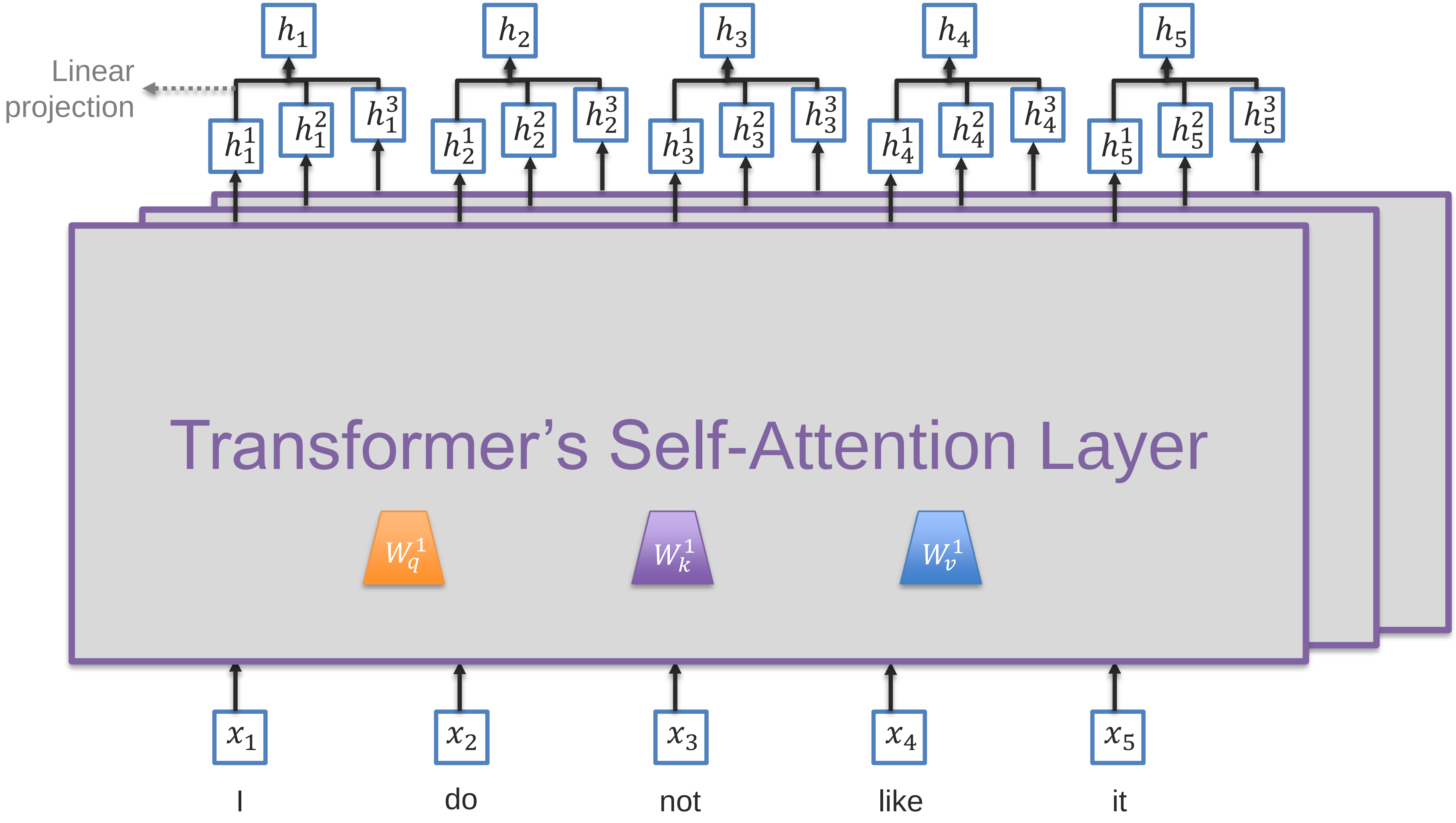


Transformer Self-Attention

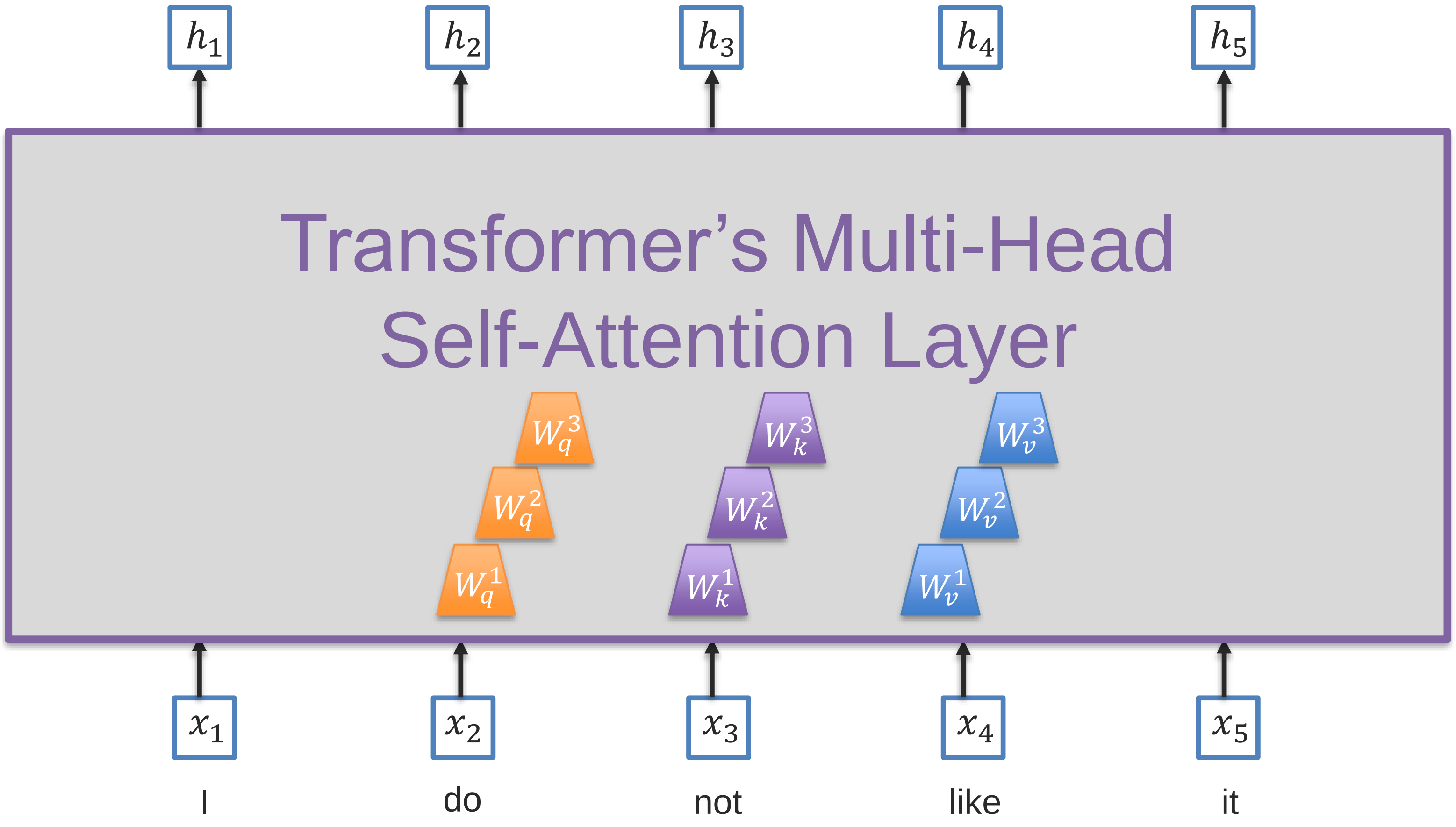
What if we want to attend simultaneously to multiple subspaces of $\boldsymbol{x}$?



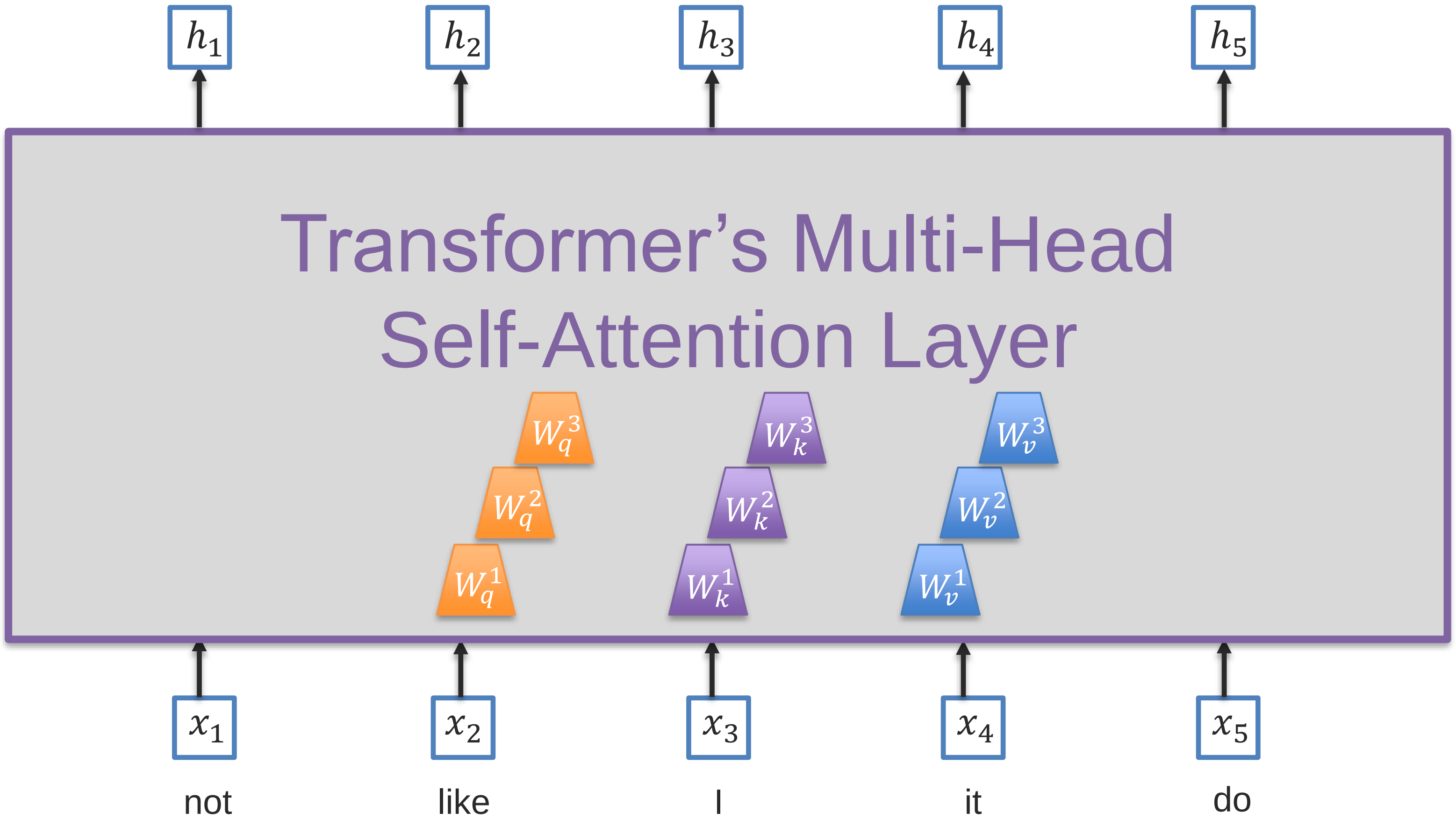
Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



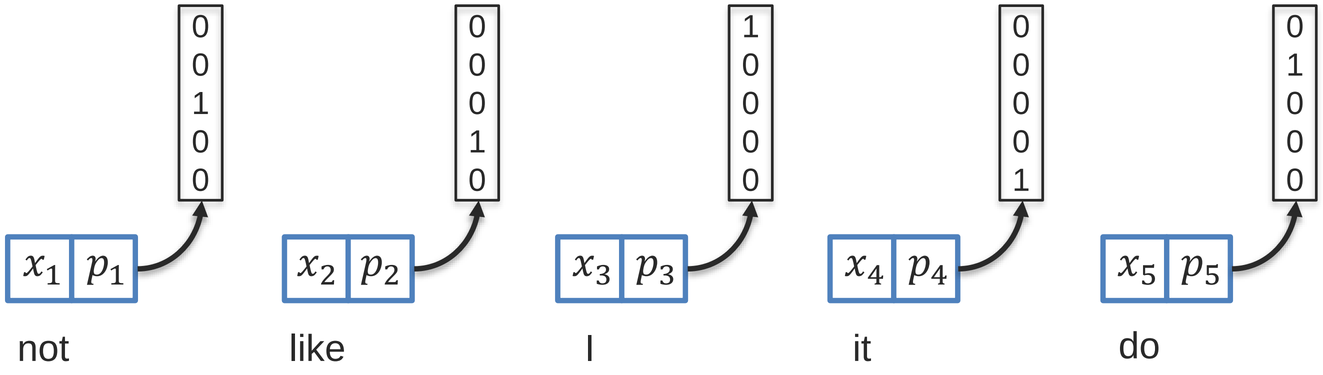
What happens if the words are shuffled?

Position embeddings

- ❑ Position information is not encoded in a self-attention module

How can we encode position information?

Simple approach: one-hot encoding

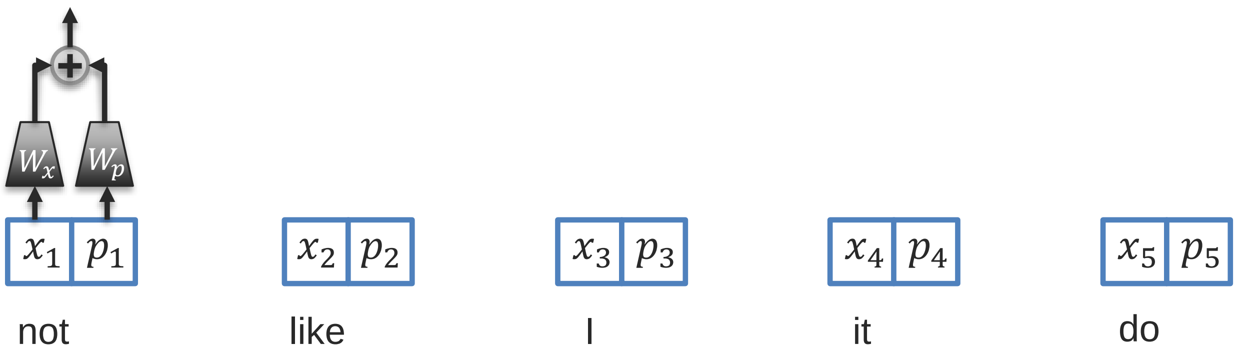


Position embeddings

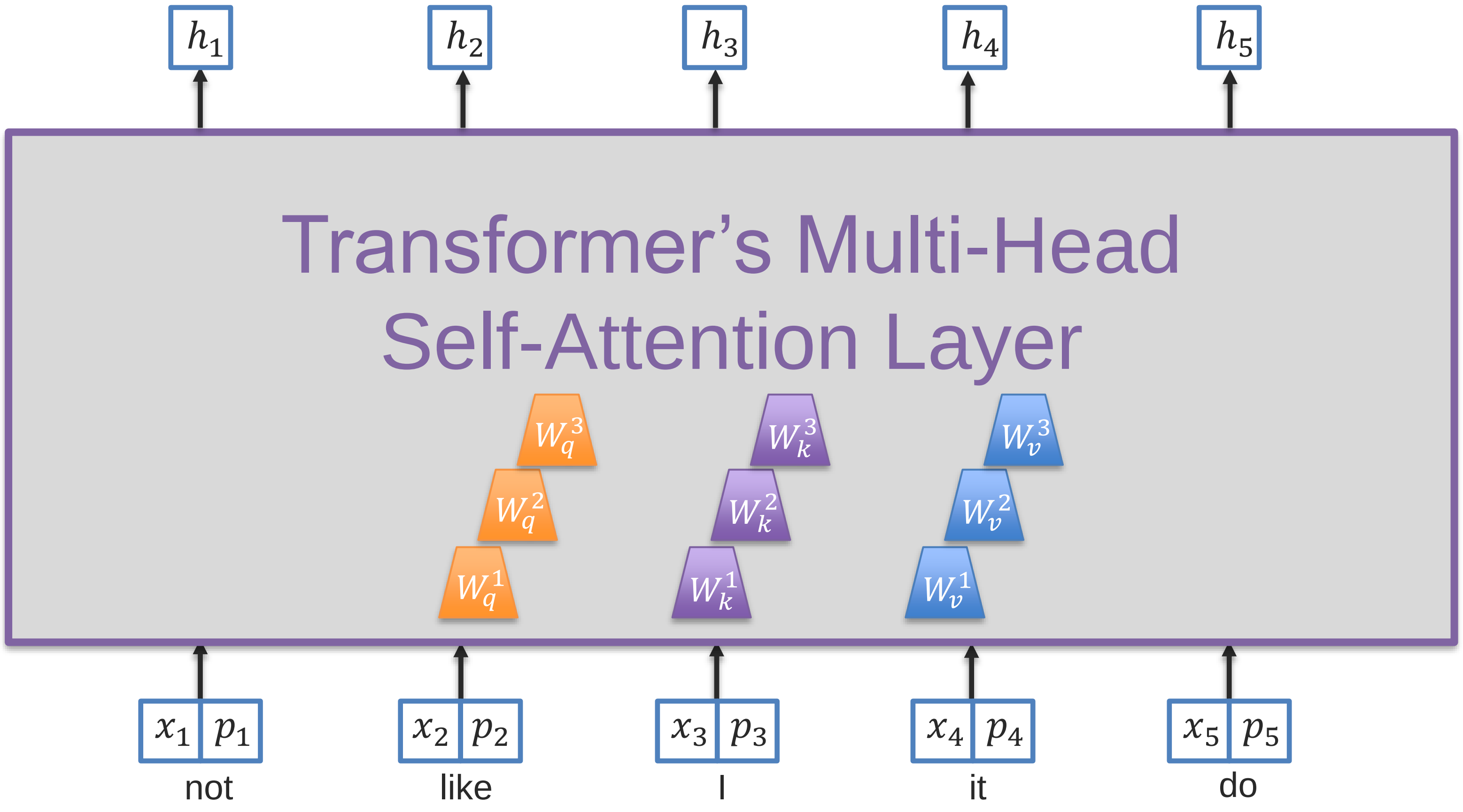
- Position information is not encoded in a self-attention module

How can we encode position information?

Simple approach: one-hot encoding + linear embeddings + $\begin{cases} \text{Sum} \\ \text{- or -} \\ \text{concat} \end{cases}$

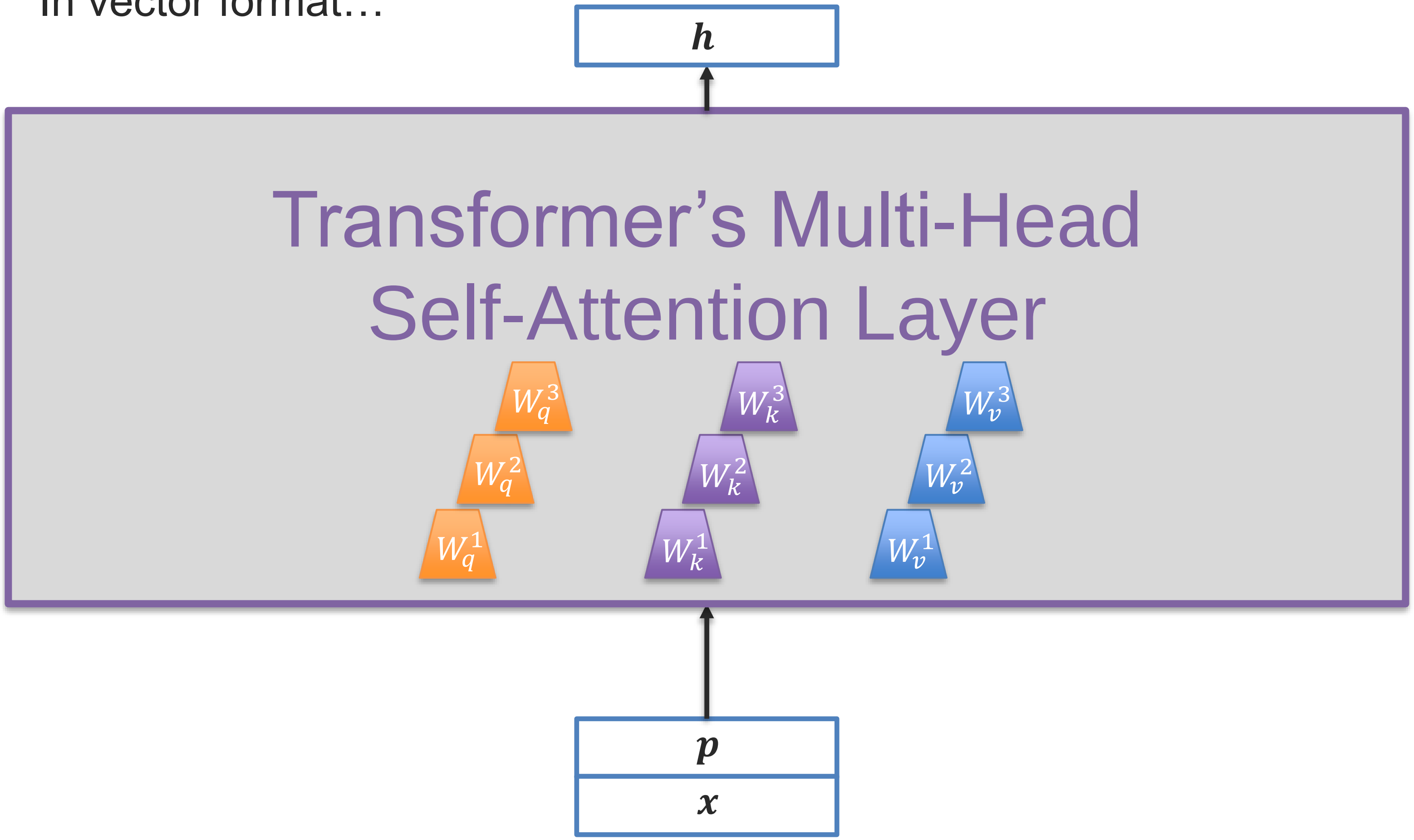


Transformer Multi-Head Self-Attention

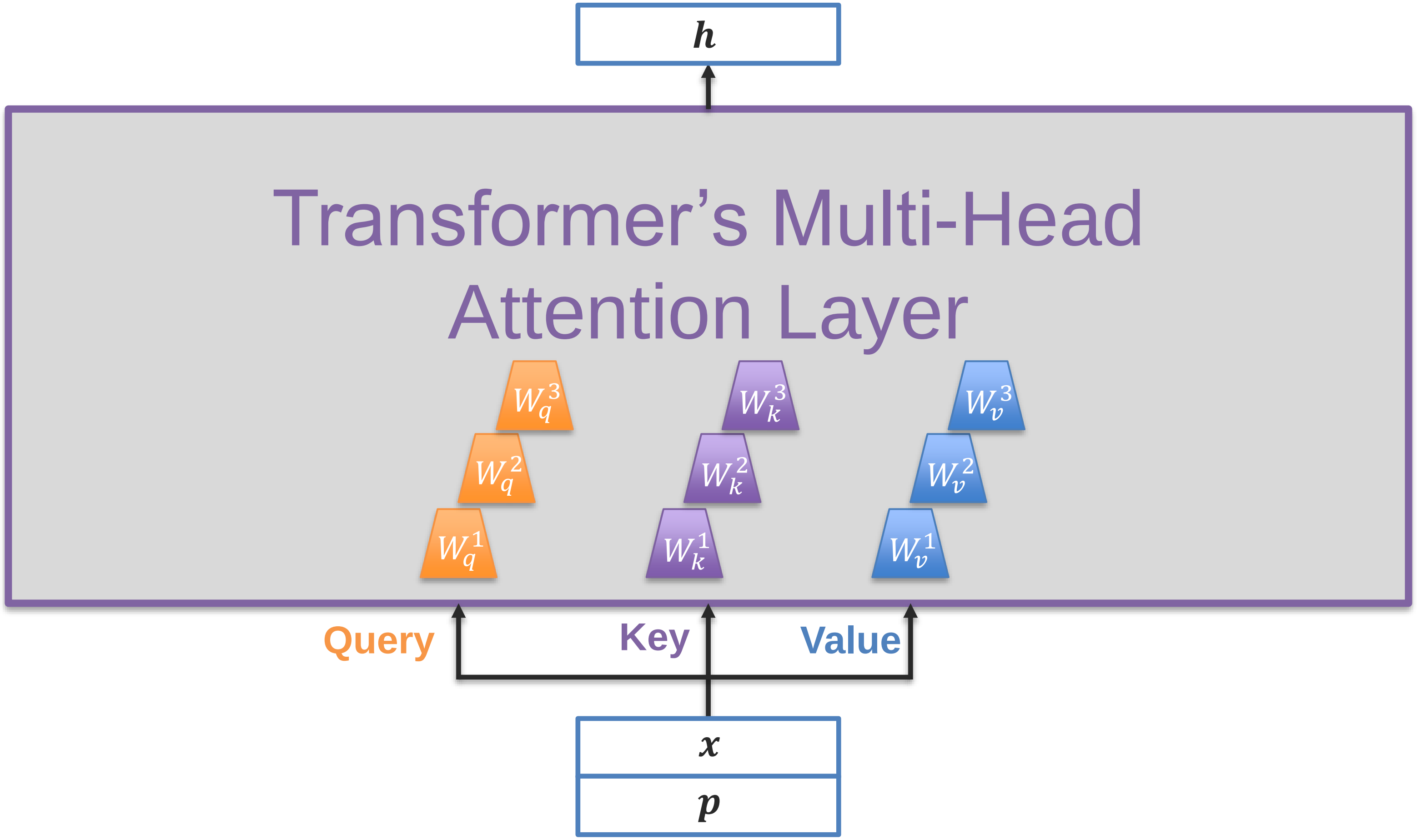


Transformer Multi-Head Self-Attention

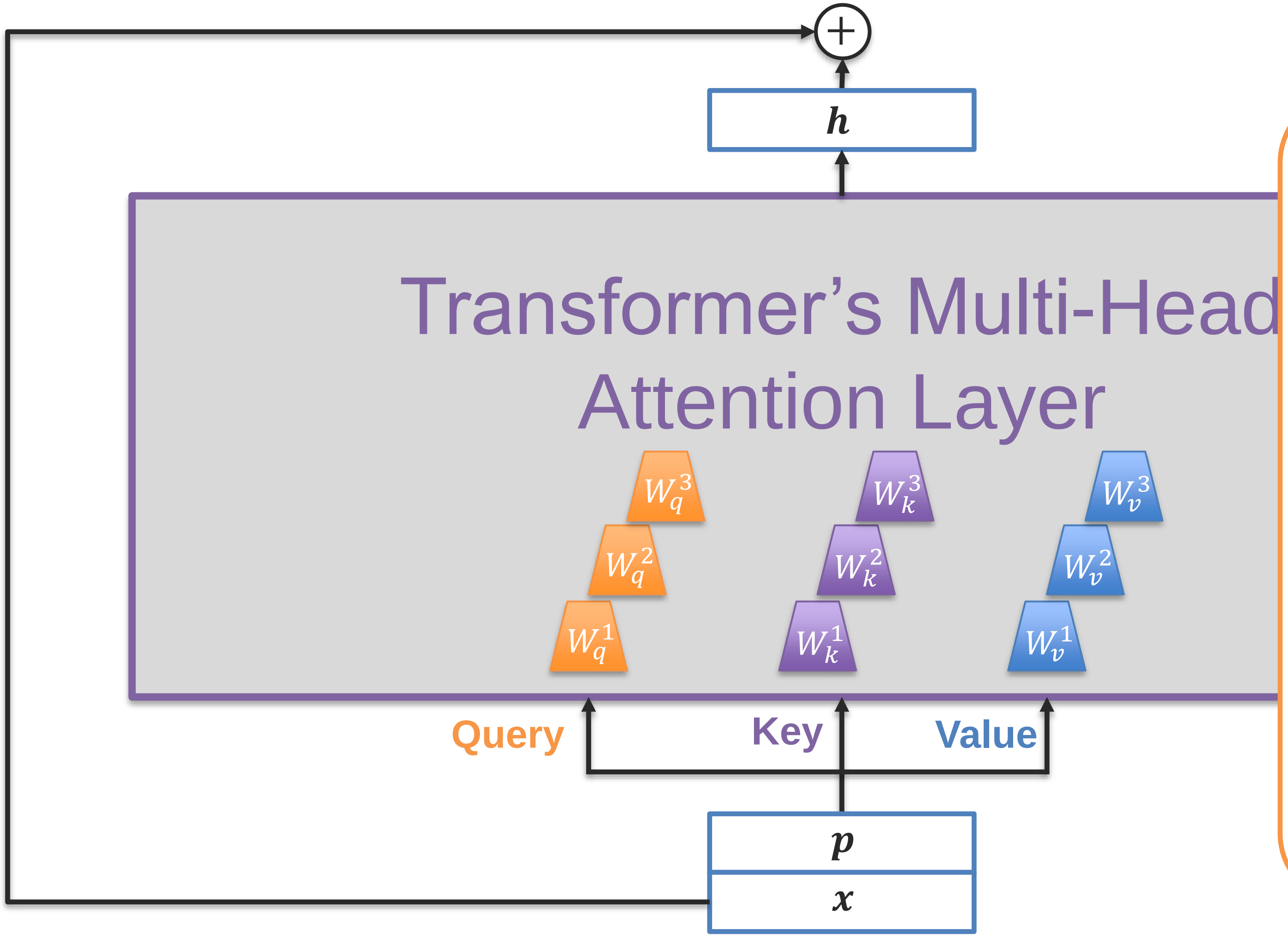
In vector format…

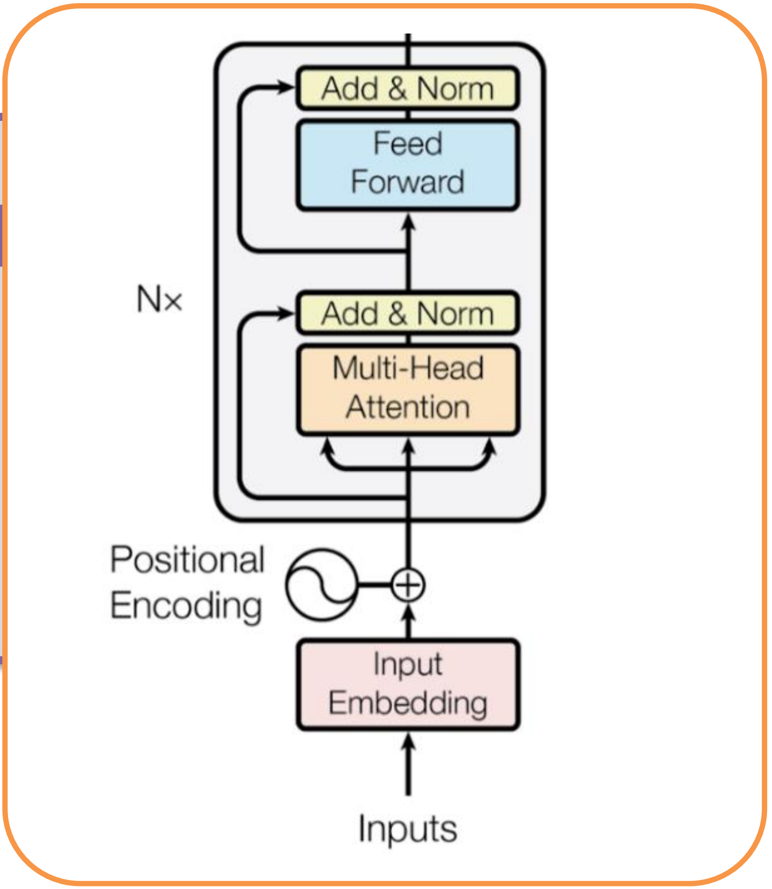


Transformer Multi-Head Attention



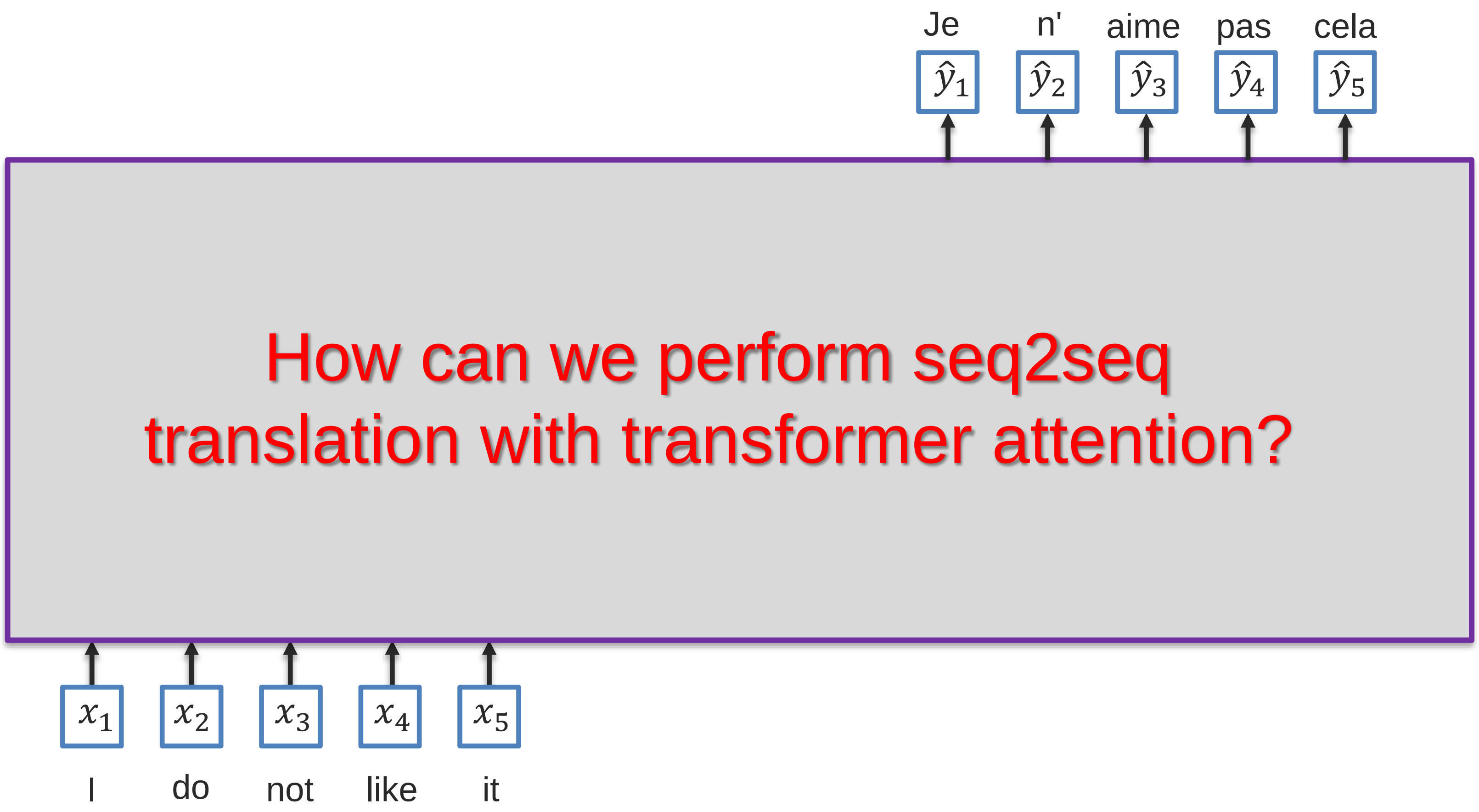
Transformer – Residual Connection



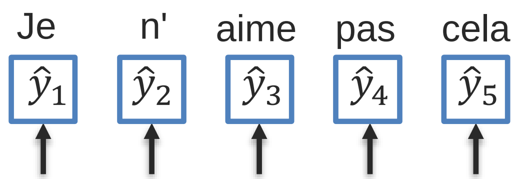


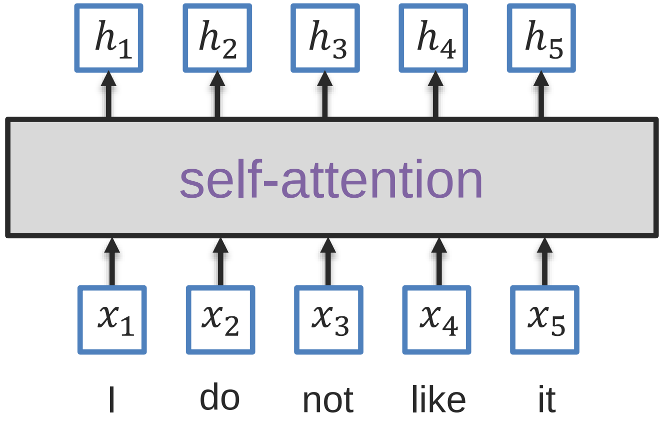
Sequence-to-Sequence Using Transformer

Sequence-to-Sequence Modeling

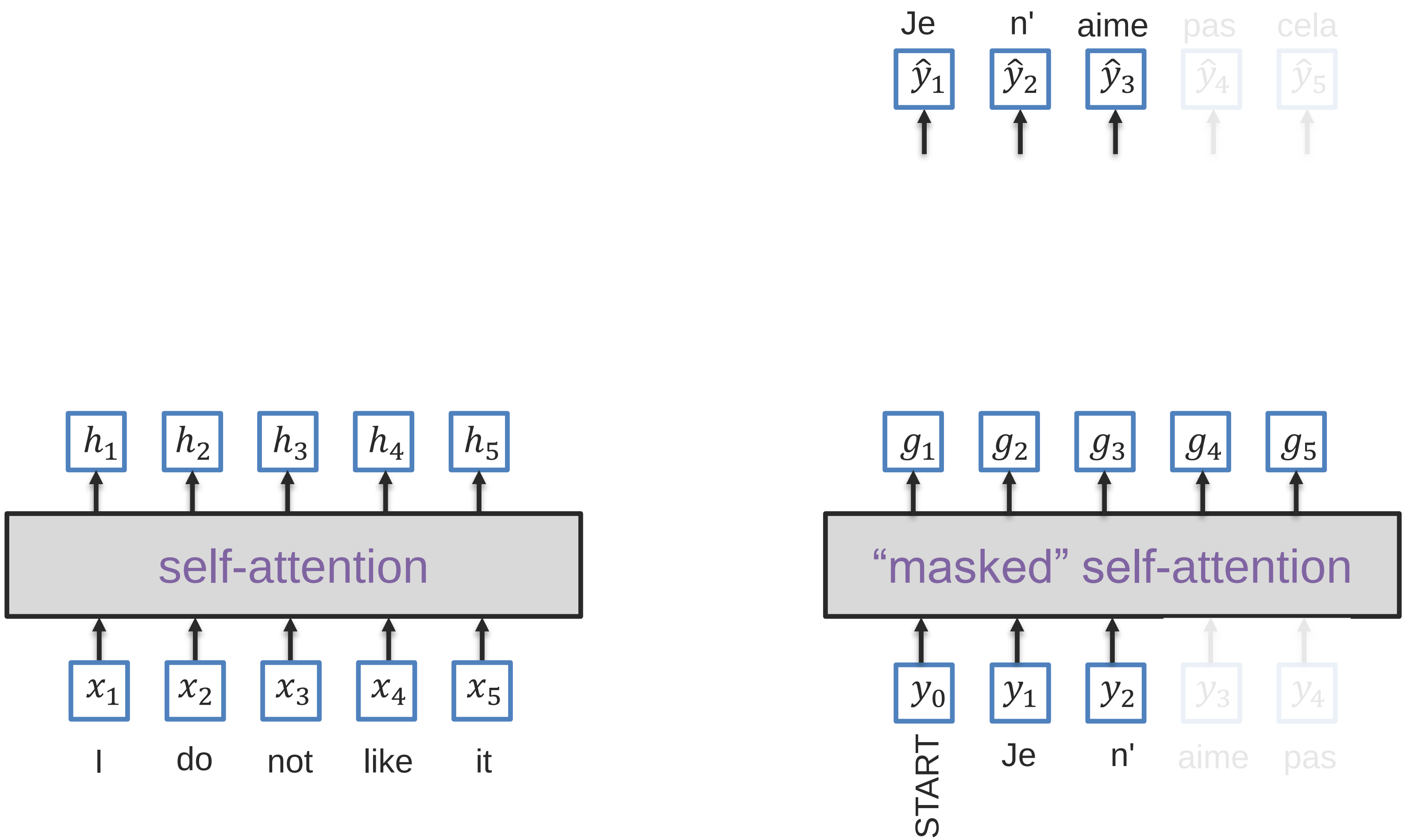


Seq2Seq with Transformer Attentions



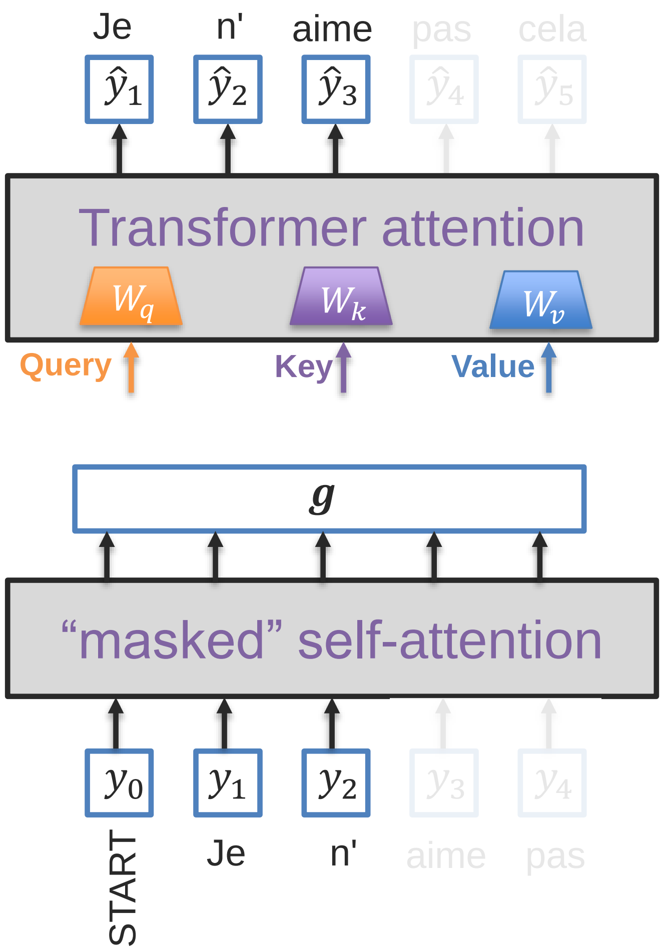


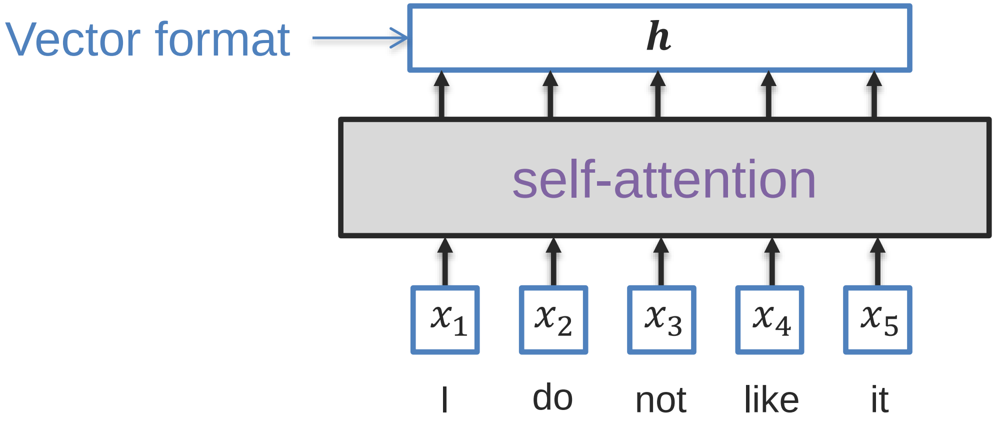
Seq2Seq with Transformer Attentions



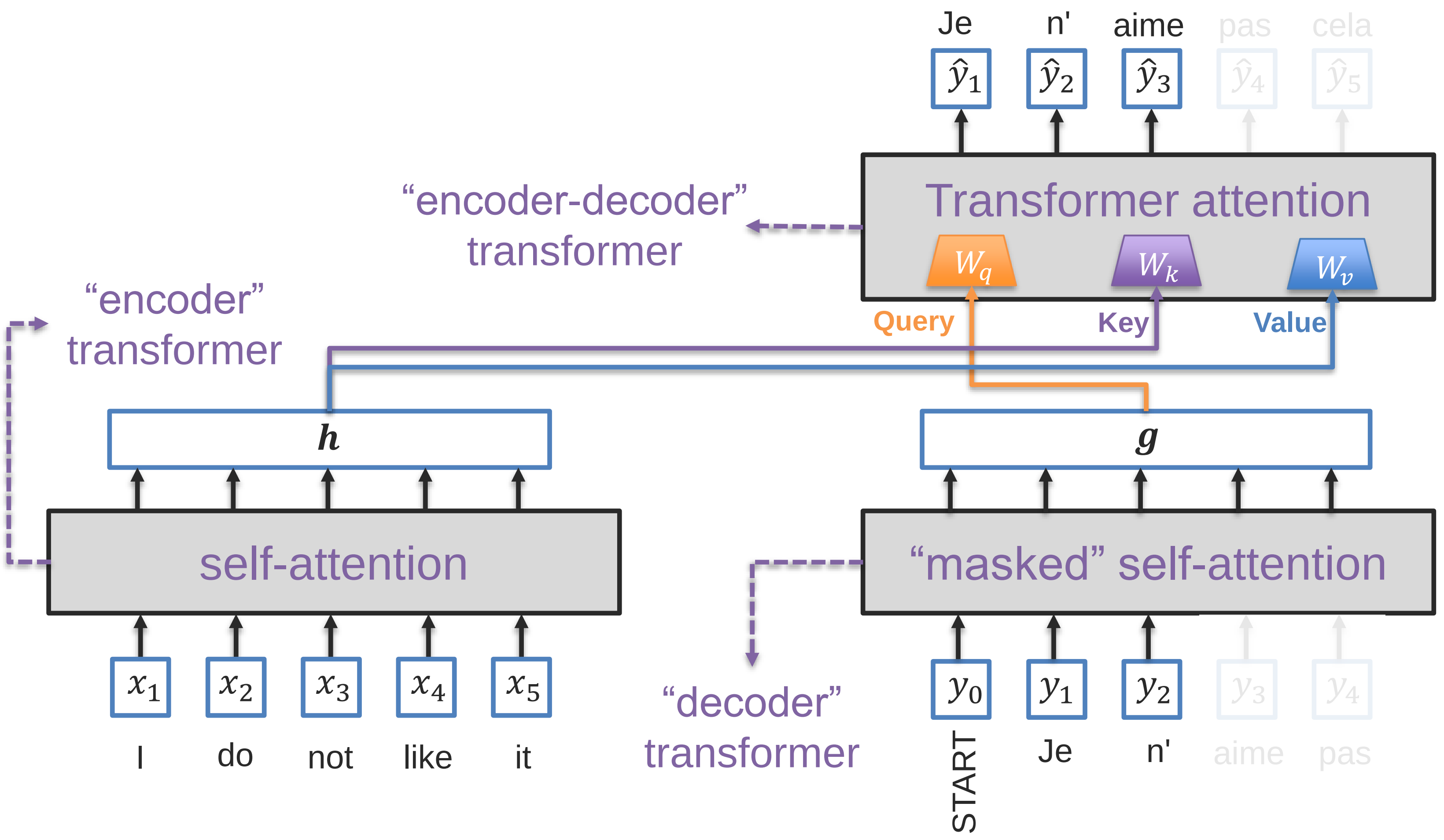
Seq2Seq with Transformer Attentions

How should we connect the encoder and decoder self-attention to the transformer attention?





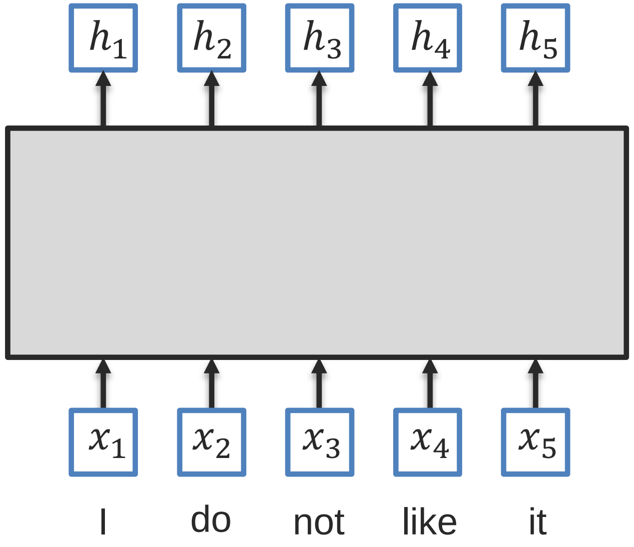
Seq2Seq with Transformer Attentions



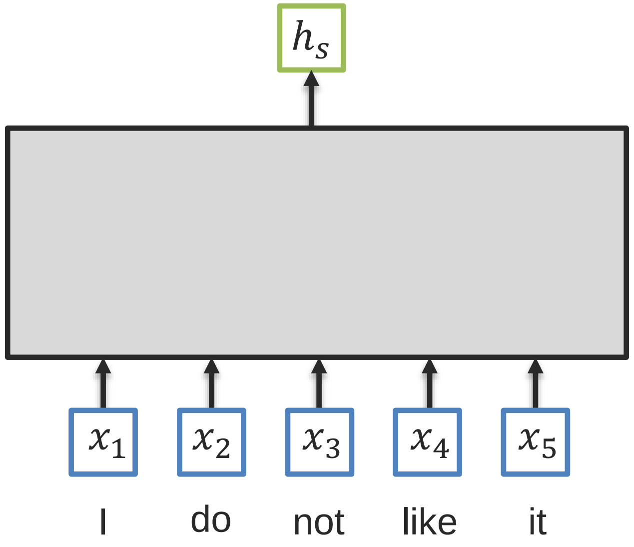
Language Pre-training

Token-level and Sentence-level Embeddings

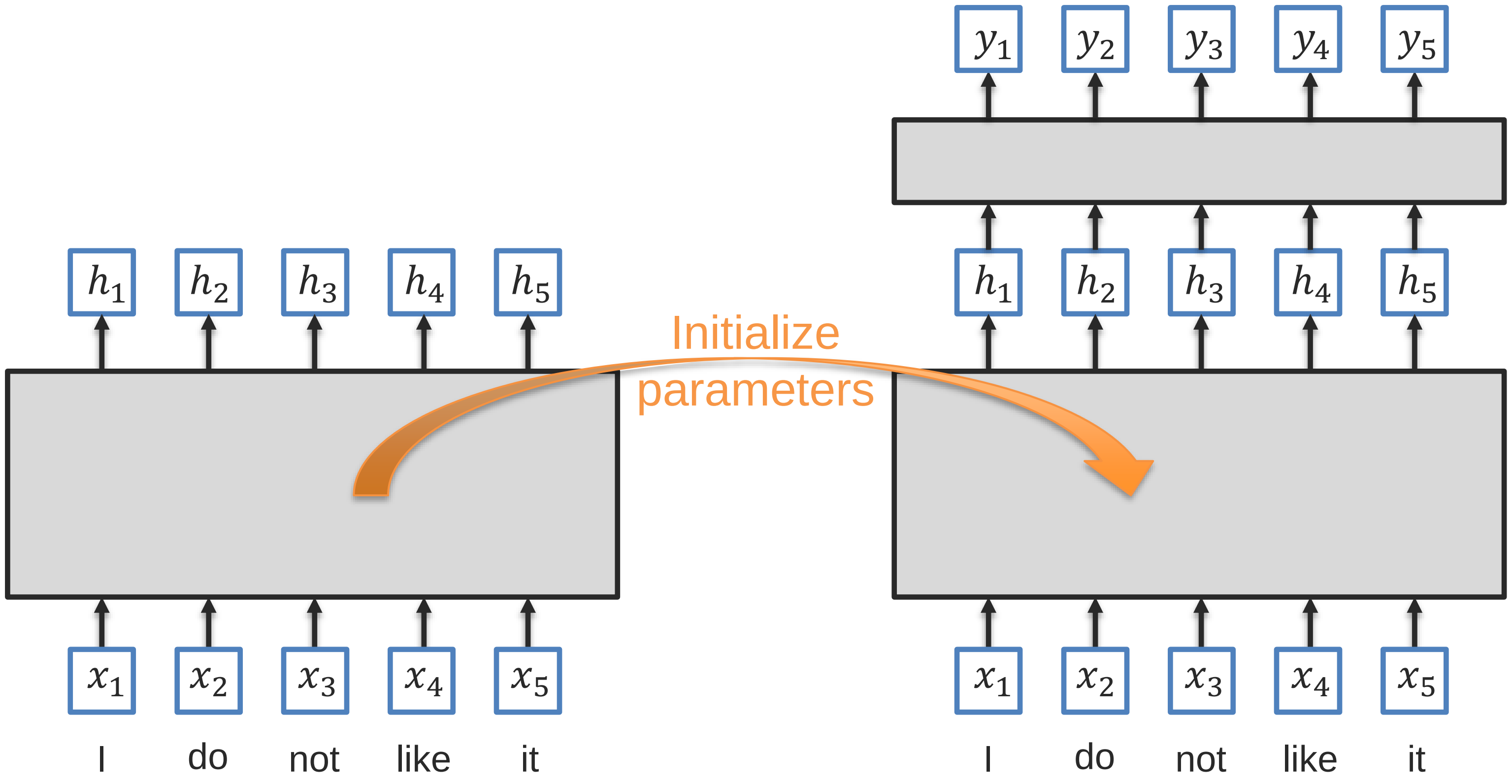
Token-level embeddings



Sentence-level embedding



Pre-Training and Fine-Tuning



Pre-training

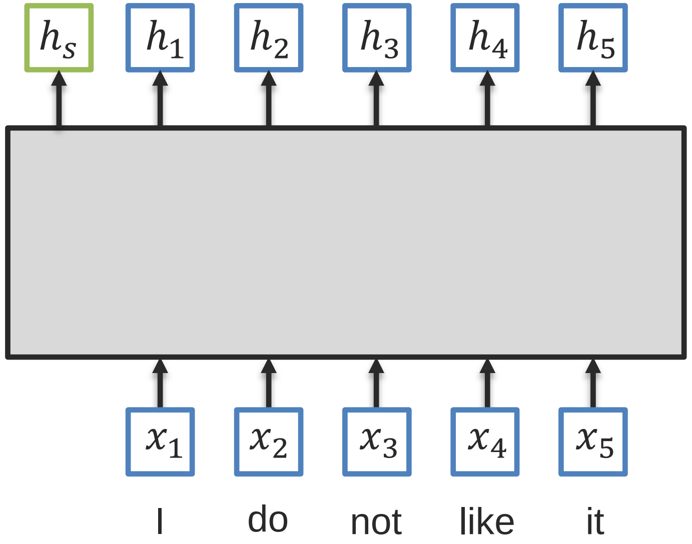
(e.g., language model)

Fine-Tuning

BERT: Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning

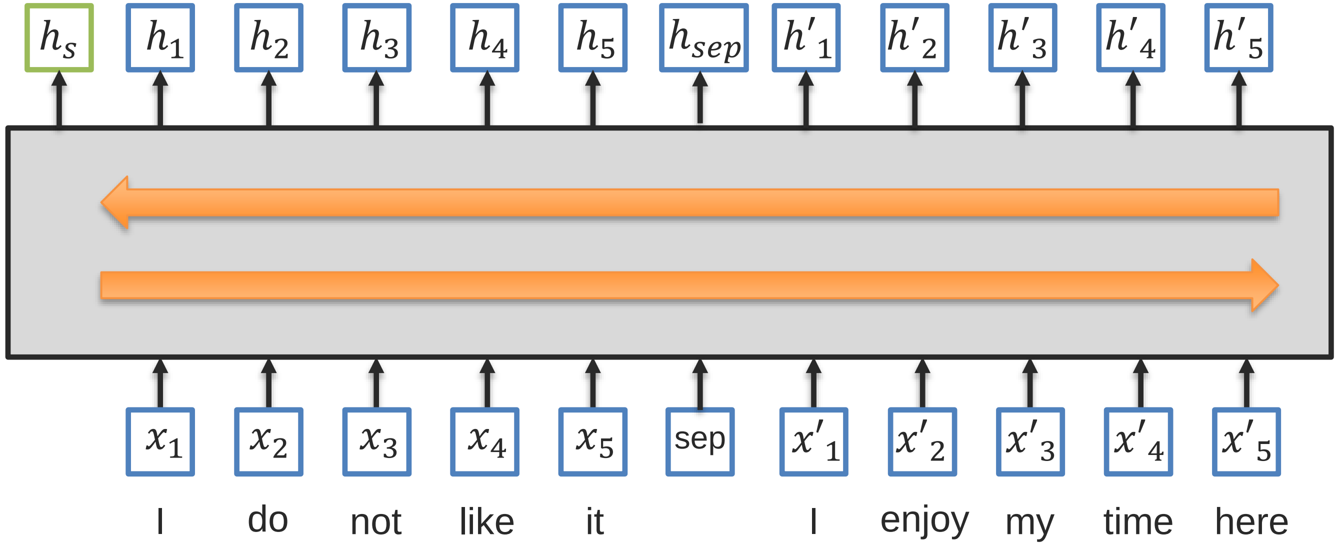


BERT: Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning
3. Can be used learn relationship between sentences
4. Models bidirectional and long-range interactions between tokens

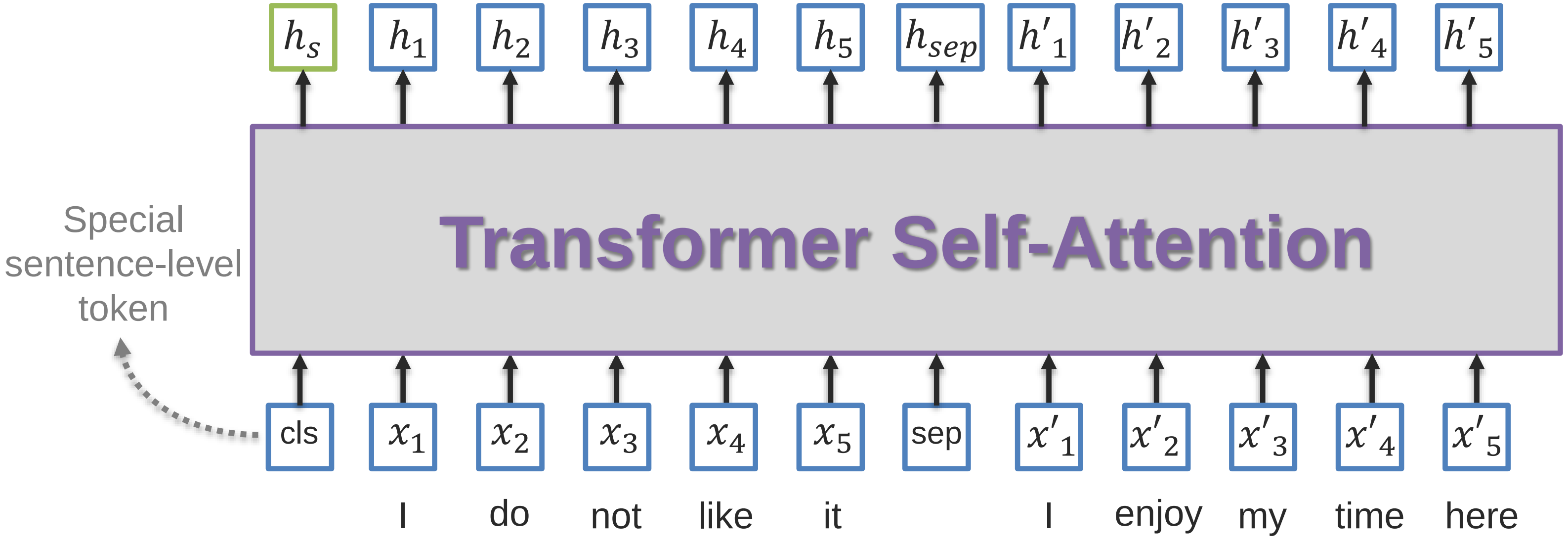
How can we do all this?



BERT: Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning
3. Can be used learn relationship between sentences
4. Models bidirectional interactions between tokens



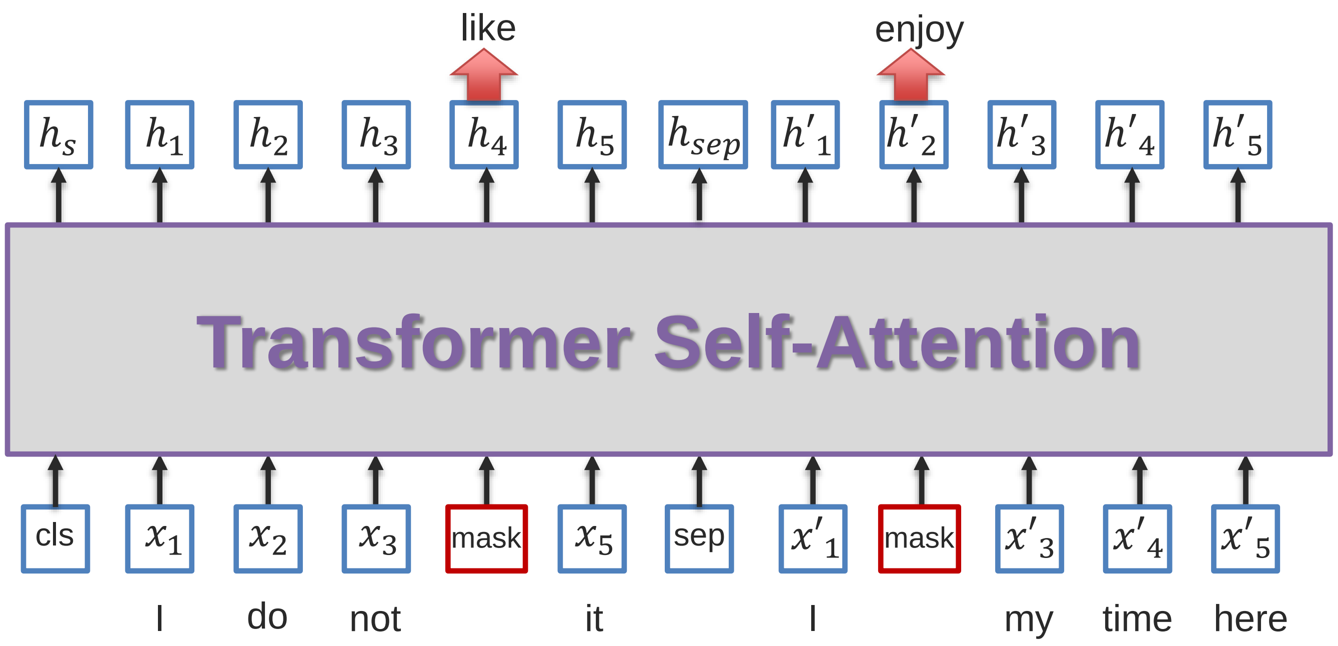
But how to train self-supervised?

Pre-training BERT Model

1 **Masked Language Model**

Randomly mask input tokens and then try to predict them

What is the loss function?



Pre-training BERT Model

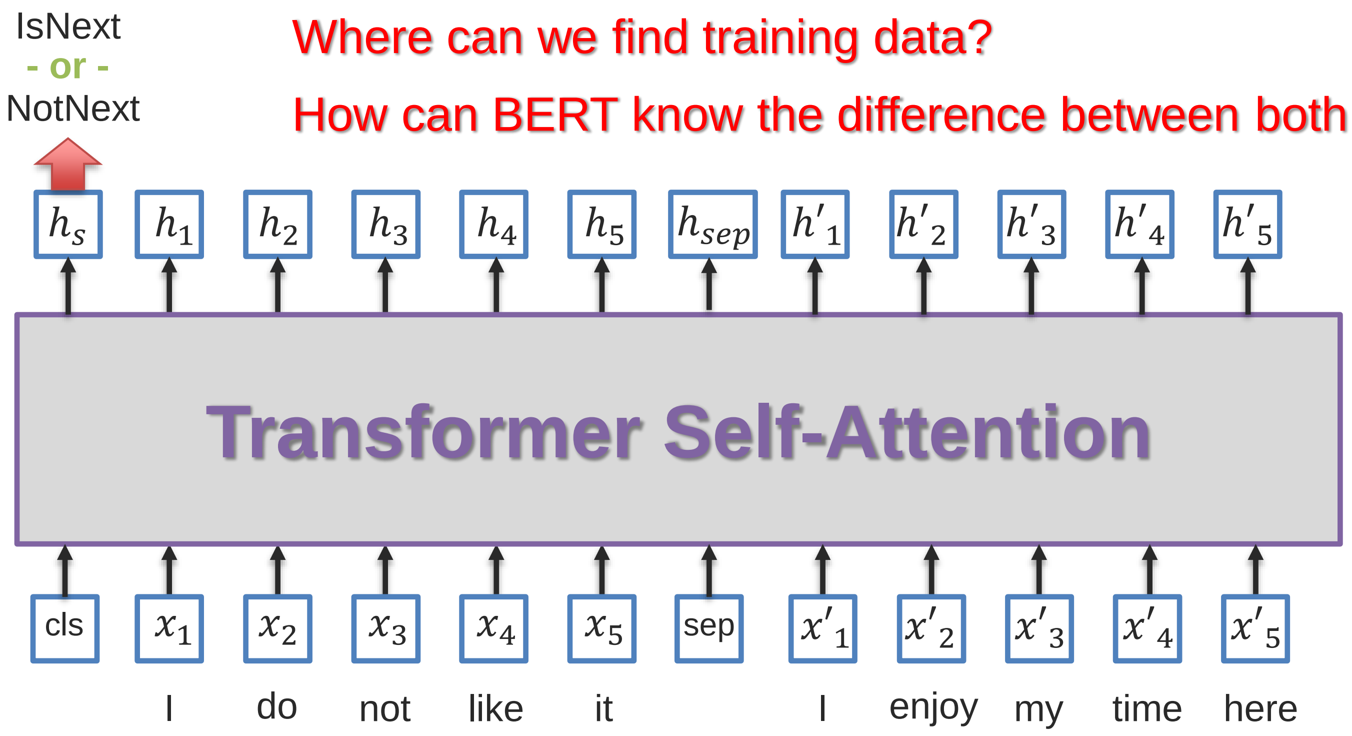
2 Next Sentence Prediction

Given two sentences, predict if this is the next one or not

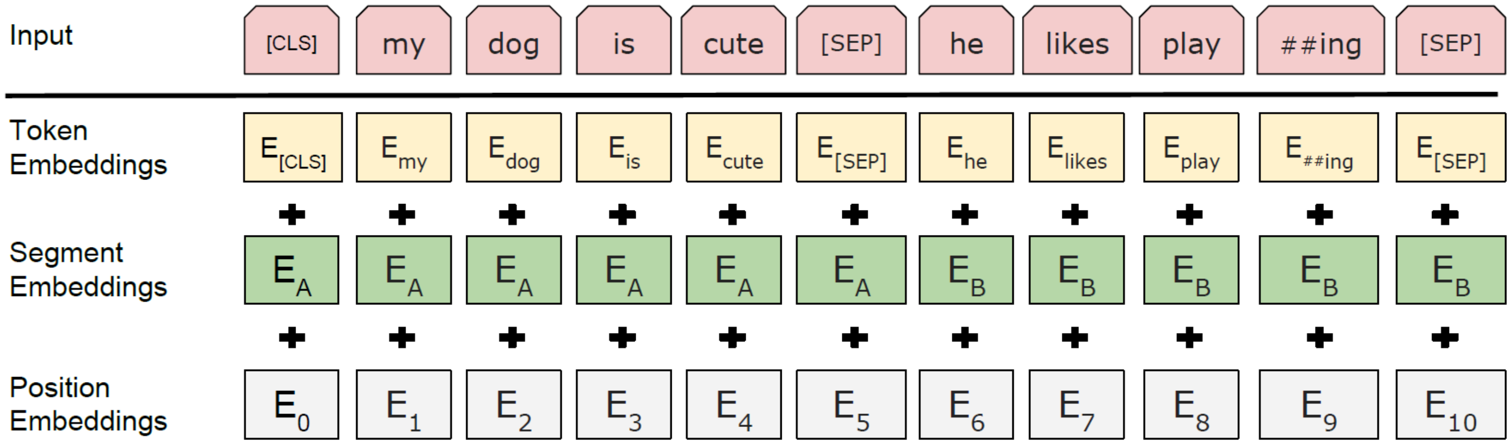
What is the loss function?

Where can we find training data?

How can BERT know the difference between both sentences?



Three Embeddings: Token + Position + Sentence

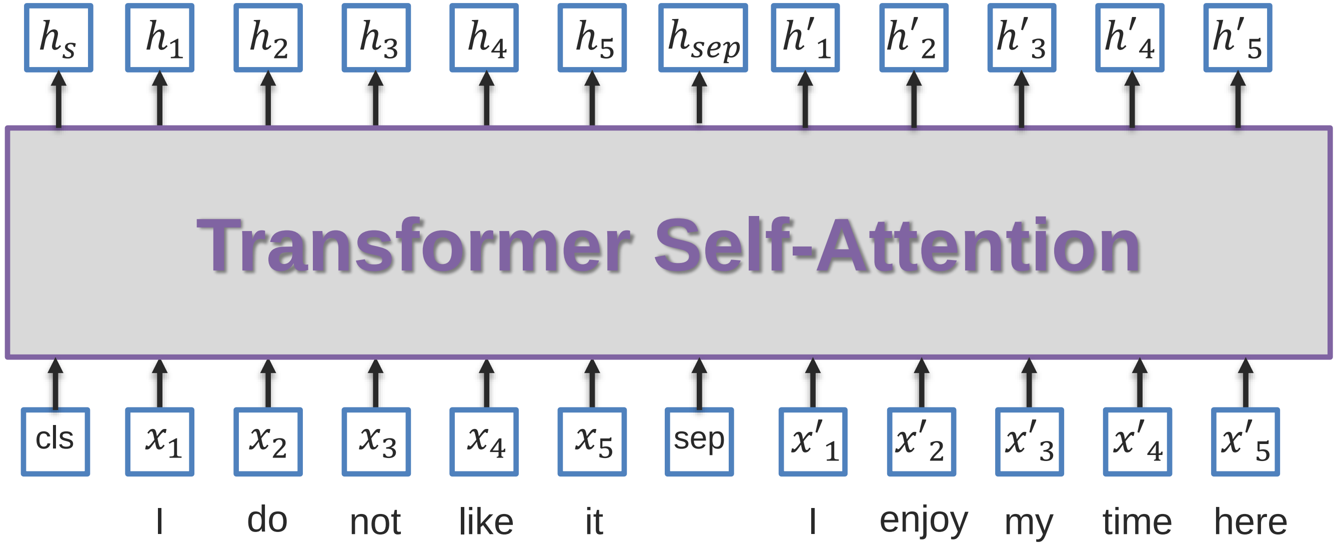


Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

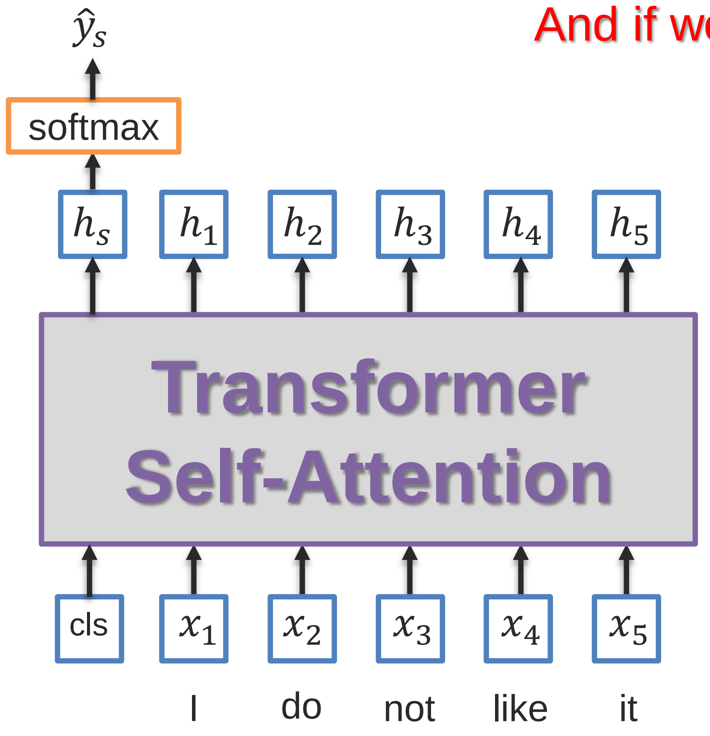
How?



Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

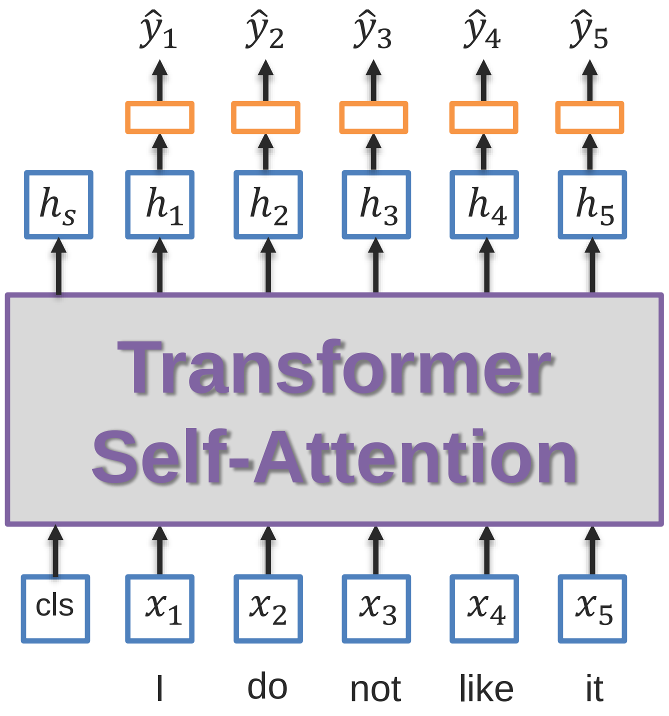


And if we have a label for each token?

Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling

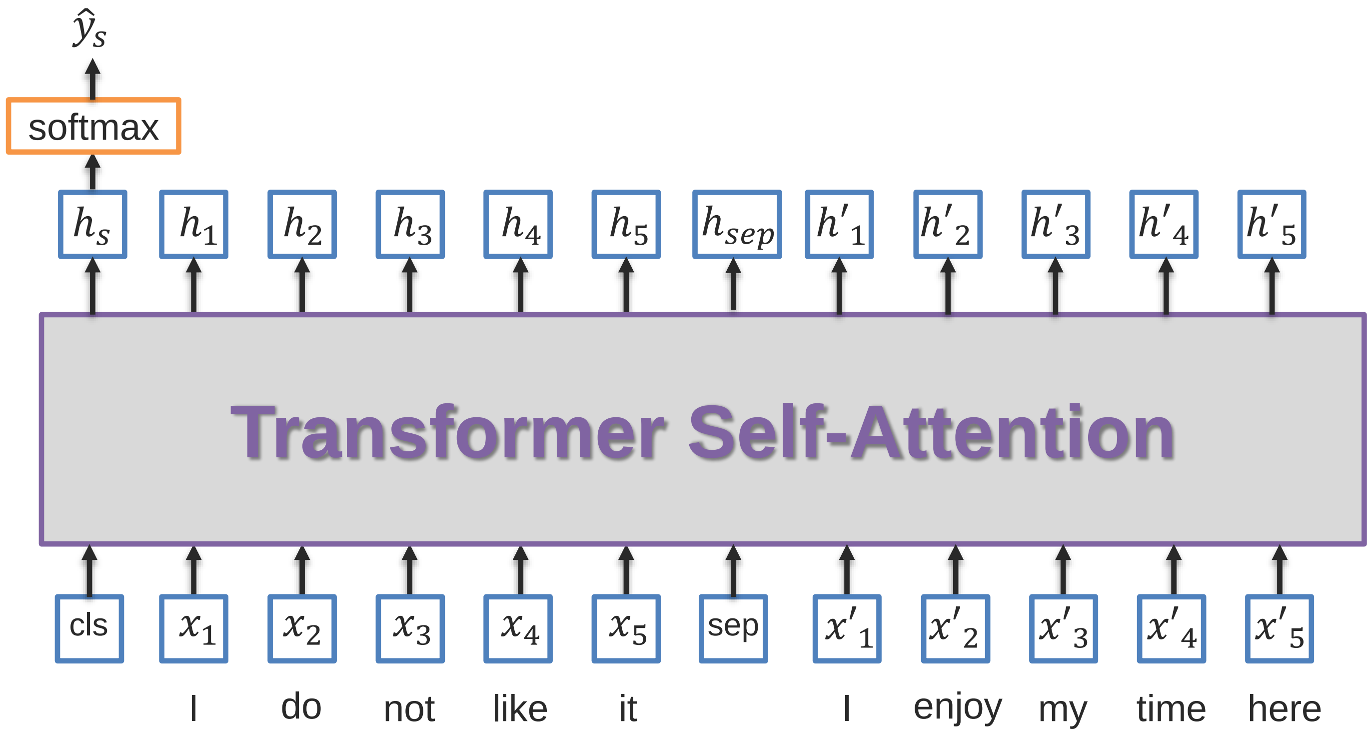


How to compare two sentences?

Fine-Tuning BERT

3 Sentence-level classification for two sentences

Examples: natural language inference



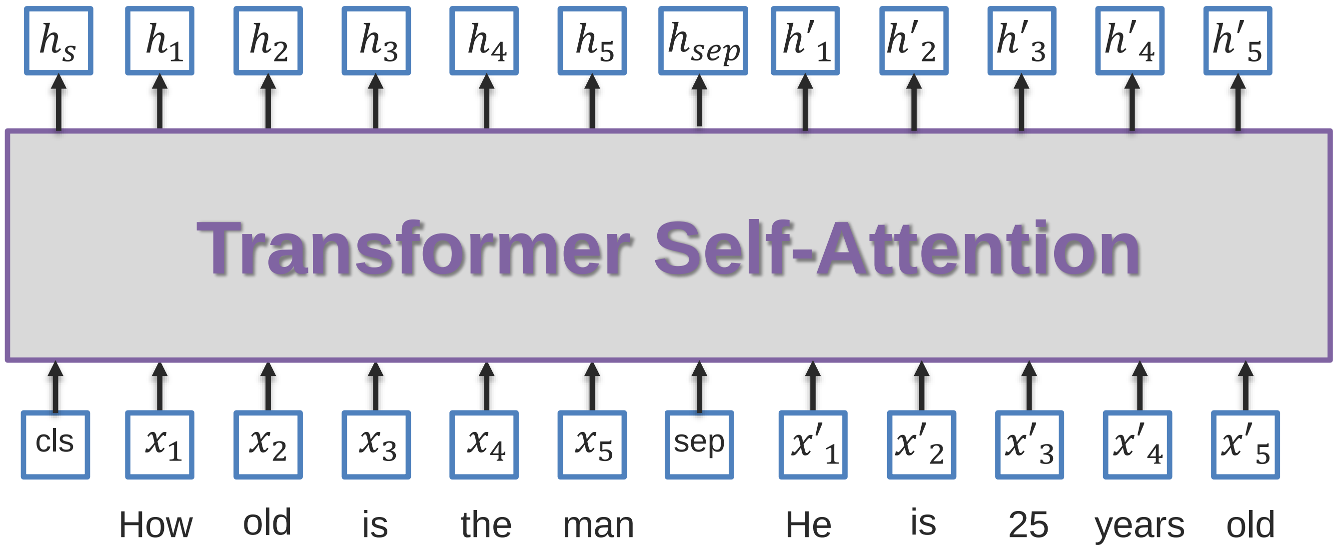
Fine-Tuning BERT

4 Question-answering: find start/end of the answer in the document

Paragraph: " … *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

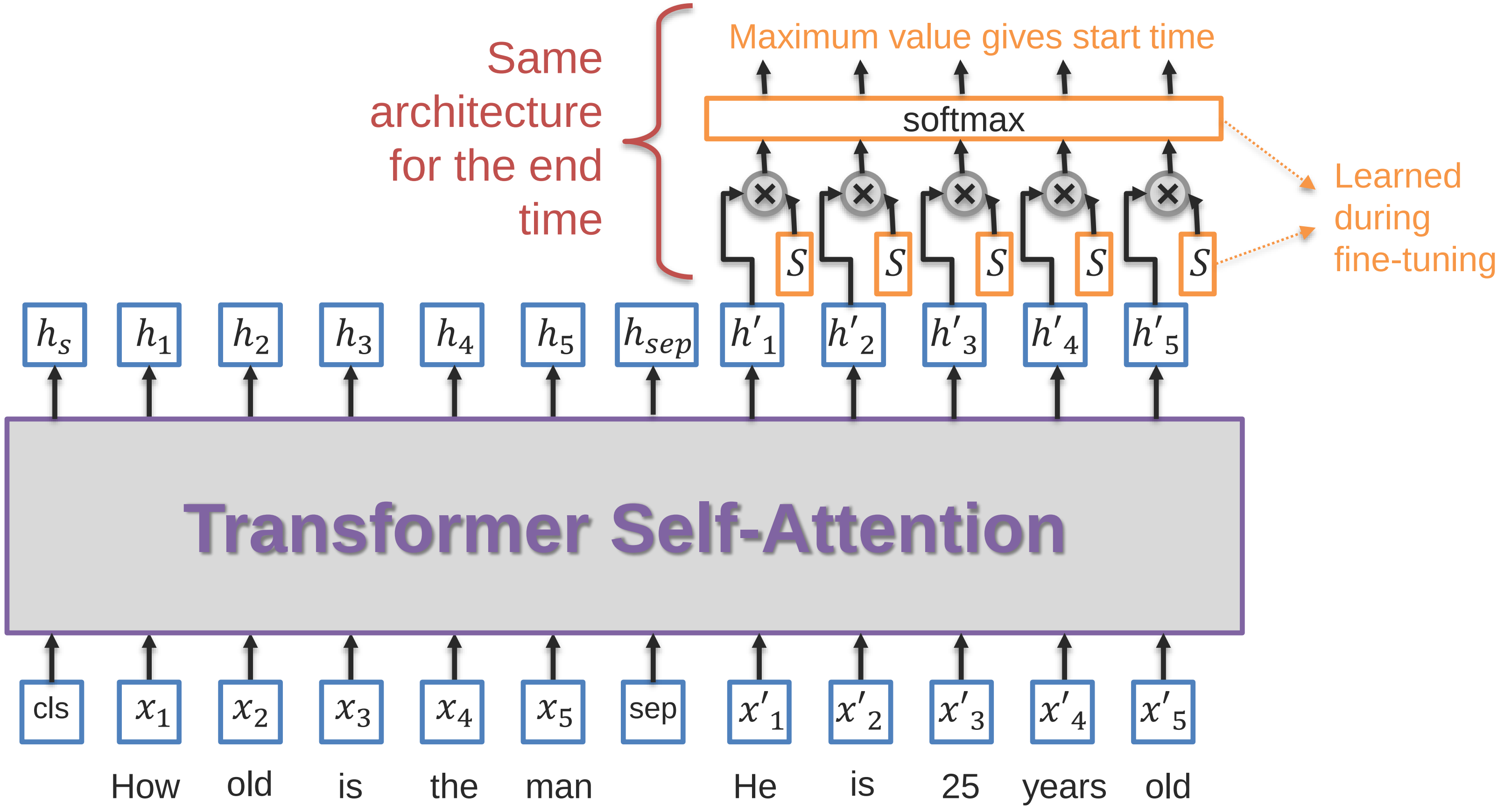
Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*



How?

Fine-Tuning BERT

4 Question-answering: find start/end of the answer in the document



And Many More…   Next week!